

2012 6th International Conference on Application of Information and Communication Technologies (AICT)

Proceedings

**Tbilisi, Georgia
17-19 October 2012**

IEEE Catalog Number: CFP1256H-ART
ISBN: 978-1-4673-1740-5



**2012 6th International Conference on Application of Information and
Communication Technologies (AICT)**

**Copyright © 2012 by the Institute of Electrical and Electronic Engineers, Inc.
All rights reserved.**

Copyright and Reprint Permissions

Abstracting is permitted with credit to the source. Libraries are permitted to photocopy beyond the limit of U.S. copyright law for private use of patrons those articles in this volume that carry a code at the bottom of the first page, provided the per-copy fee indicated in the code is paid through Copyright Clearance Center, 222 Rosewood Drive, Danvers, MA 01923.

For other copying, reprint or republication permission, write to IEEE Copyrights Manager, IEEE Service Center, 445 Hoes Lane, Piscataway, NJ 08854. All rights reserved.

IEEE Catalog Number: CFP1256H-ART
ISBN: 978-1-4673-1740-5

Printed copies of this publication are available from:

Curran Associates, Inc
57 Morehouse Lane
Red Hook, NY 12571 USA
Phone: (845) 758-0400
Fax: (845) 758-2633
E-mail: curran@proceedings.com

Produced by IEEE eXpress Conference Publishing

For information on producing a conference proceedings and receiving an estimate, contact conferencepublishing@ieee.org
<http://www.ieee.org/conferencepublishing>

CONFERENCE COMMITTEES

CONFERENCE HONORARY CHAIR

Nodar Surguladze

Deputy Minister of Education and Science of Georgia

CONFERENCE CHAIRS

Prof. Ahmet Sanich,

Qafqaz University Rector, Azerbaijan

Prof. Ercan Tunc,

International Black Sea University, Georgia

Prof. Alexander Kvitashvili,

Tbilisi State University Rector, Georgia

GENERAL CHAIR

Assoc. Prof. Abzetdin Adamov, *Qafqaz University, Azerbaijan*

TECHNICAL PROGRAM CO-CHAIRS

Prof. H.Levent Akin,

Bogazici University, Turkey

Associate Prof. Vincent Guyot,

ESIEA/LIP6, France

Prof. Asoke Talukder,

IIT, Bangalore, India

Prof. Cevdet Meriç,

Fatih University, Turkey

Prof. Serdar Korukoğlu,

Ege University, Turkey

Assistant Prof. Kshetrimayum Rakhesh Singh,

IIT-Guwahati, Guwahati, Assam, India

Dr. Nazim Agoulmine,

University of Evry Val d'Essonne, France

Dr. Nargiza Usmanova,

IEEE Comsoc Chapter chair, TUIT, Uzbekistan

Prof. Alok Kumar Das,

Guru Nanak Institute of Technology, India

PANEL CHAIR

Assoc. Prof. Abzetdin Adamov, *Qafqaz University, Azerbaijan*

TUTORIALS AND PUBLICATION CHAIR

Assoc. Prof. Abzetdin Adamov, *Chair of Computer Engineering Department at the Qafqaz University*

LOCAL ORGANIZING COMMITTEE

Assoc.Prof, Cabir Erguven

Dean of Faculty of Engineering, IBSU

Dr. Ghvedashvili Giorgi,

Associate Professor, Head of Department of Scientific Research and Development, TSU

Prof. Bochorishvili,

Dean of Faculty of Exact and Natural Sciences, TSU

Prof. Alexander Gamkrelidze,

Department of Computer Sciences, TSU

Prof. Koba Gelashvili,

Department of Computer Sciences, TSU

Prof. Gia Sirbiladze,

Department of Computer Sciences, TSU

Prof. Manana Khachidze,

Department of Computer Sciences, TSU

Dr. Cihan Mert,
Head of Master Programs, IBSU
Dr. Vakhtang Rodonaia,
Vice Dean, IBSU
Dr. Akinnola Akintunde,
Head of PhD Programs and Scientific Researches, IBSU
Okan Eray,
Lecturer, IBSU

ADVISORY AND STEERING COMMITTEE

Niftali Gocayev,
Vice Rector for Research and Science, Qafqaz University, Azerbaijan
Victor Kureichik,
Taganrog State University, Russia
Mehmet Ulema,
Manhattan College, USA
Giuli Alasanai,
Prof. Dr., Vice Rector, IBSU
Raouf Boutaba,
University of Waterloo, Canada
Adnan Yazici,
Middle East Technical University, Turkey
Guy Omidyar,
IFIPTC6 WG6.8 Past Chair, USA

TECHNICAL PROGRAM COMMITTEE MEMBERS

Cem Ersoy,	<i>Bogazici University, IEEE ComSoc Chapter Chair, Turkey</i>
Abolfazl Mehbodniya,	INRS, Canada
Asoke K Talukder,	IIT Bangalore, India
Babek Abbasov,	Qafqaz University, Azerbaijan
Etibar Seyidzade,	Qafqaz University, Azerbaijan
Canchi Radhakrishna,	KYOCERA, USA
Djamel Sadok,	Universidade Federal de Pernambuco, Brazil
Gholam Ali Rezai rad,	University of Science and Technology, Iran
Guy Omidyar,	Consultant, USA
Agasi Melikov,	National Aviation Academy, Azerbaijan
Nadir Alishov,	Ukrain National Academy of Science, Ukrain
Guy Pujolle,	Université Paris 6, France
Alekber Aliyev,	Baku State University, Azerbaijan
Ramin Mahmudzade,	Baku State University, Azerbaijan
Petr Sosnin,	Ulyanovsk State Technical University, Russia
Khaldoun Al Agha,	LRI, University of Paris XI, France
Khaled Ben Letaief,	Hong Kong University, HK
Halil Ismailov,	Qafqaz University, Azerbaijan
Marie-Jose Montpetit,	Motorola, USA
Mehmet Ulema,	Manhattan College, USA
Masoud Mashhour,	Carleton University, Canada
Tugrul YANIK,	Fatih University, Turkey
Fatih CAMCI,	Fatih University, Turkey
Nazar Elfadil,	SQU, Oman
Ruslan Kamaev,	KSUCTA, Kyrgyzstan
Abdul Halim Zaimi,	Istanbul University, Turkey
Kamil Saraç,	University of Texas at Dallas, USA
Hasan Bulut,	Ege University, Turkey
Muhammet Cinsdikici,	Ege University, Turkey
Chandana P. Withana,	Charles Sturt University, Australia
Ahmet Burak Can,	Hacettepe University, Turkey

TABLE OF CONTENTS

Ontology Creation for an Educational Center	3
<i>Selma Dilek, Hacer Karacan, Nesa Jahangiri and Shima Afzali</i>	
New novel idea for Cloud Computing: How can we use Kalman filter in security of Cloud Computing	7
<i>Mehdi Darbandi, Pariya Shahbazi, Saeed Setayesh and Ole-Christoffer Granmo</i>	
Particle Swarm Intelligence as a New Heuristic for the Optimization of Distributed Database Queries	12
<i>Tansel Dokeroglu, Umut Tosun and Ahmet Cosar</i>	
Realization of Direct Series of Discussions in the Face of Conflicting Set of Production Rules	19
<i>Roman Samkharadze, Davit Chikovani and Lia Gachechiladze</i>	
Towards Remote Security Monitoring in Cloud Services Utilizing Security Metrics	23
<i>Reijo M. Savola and Jukka Ahola</i>	
Real Time Decision Support Systems for Mobile Users in Intelligent Cities	30
<i>Alfio Costanzo and Alberto Faro</i>	
Lip reading using fuzzy logic network with memory	35
<i>Stefan Badura, Martin Klimo and Ondrej Skvarek</i>	
Analysis of a Novel Audio Hash Function Based upon Stationary Wavelet Transform	39
<i>Mahdi Nouri, Zahra Zeinolabedini, Nooshin Farhangian and Nasim Fekri</i>	
Application of geometrical algebra to neural computations	45
<i>Irakli Rodonaia and Vakhtang Rodonaia</i>	
The adaptive method of decision making in problems of motion terminal control	49
<i>Vakhtang Rodonaia</i>	
BSS parameters and their influences in GSM mobile networks	53
<i>Mohammad Reza Salehifar, Saeed Soleimany and Hassan Karbalaee</i>	
Feature Based Iris Recognition System Functioning on Extraction of 2D Features	57
<i>Arjun Agrawal, Gundeep Singh Bindra and Priyanka Sharma</i>	
Structuring of Normative-Technical Documents in Information and Computing Environment Taking into Account Ageing of Information	62
<i>V.G. Lim, I.G. Voevodin, N.B. Tashpulatova and B.T. Kabulov</i>	
Labeled Protection of Project Tasks in Collaborative Designing of Software Intensive Systems	64
<i>P. Sosnin, V. Maklaev and S. Zhukov</i>	
Integrated Analytical Information Resource Management System	69
<i>Giorgi Ghlonti</i>	

NFC: Smart Recording Of Traffic Violation System	71
<i>Omid Nejati and Mohsen Yaghoubi Suraki</i>	
Rule Scheduling Methods in Active Database Systems: A Brief Survey	75
<i>Hamid-Reza Firoozy-Najafabadi and Ahmad Habibizad Navin</i>	
Analysis of Inbound and Outbound Email Traffic and its SPAM Impact	80
<i>Seema Khanna and Harish Chaudhry</i>	
Research on Evaluation Techniques for Immersive Multimedia	88
<i>Aslinda Md. Hashim, Fakaruddin Fahmi Romli and Zosipha Zainal Osman</i>	
Collision Avatar (CA): Adding Collision Objects for Human Body in Augmented Reality using Kinect.....	95
<i>Kairat Aitpayev and Jaafar Gaber</i>	
A Comparative Study on Feature Selection in Chinese Spam Filtering	99
<i>Yan Xu</i>	
Web Content Reauthoring for Large Screens	105
<i>Neetu Narwal and Saba Hilal</i>	
Feature Based Iris Recognition System Functioning on Extraction of 2D Features.....	108
<i>Arjun Agrawal, Gundeep Singh Bindra and Priyanka Sharma</i>	
The Bus Arrival Time Service Based on Dynamic Traffic Information	113
<i>Tongyu Zhu, Jian Dong, Jian Huang, Songsong Pang and BoWen Du</i>	
Application Test Process in Product Life Cycle.....	119
<i>Oner Tekin and Gulsah Bayram Cetin</i>	
Distributed File System as a basis of Data-Intensive Computing.....	125
<i>Abzetdin Adamov</i>	
Direct Robust Non-Negative Matrix Factorization and Its Application on Image Processing.....	128
<i>Bin Shen, Zhanibek Datbayev and Olzhas Makhambetov</i>	
Electronic Health Card: Opportunities and Challenges	133
<i>Hamid-Reza Firoozy-Najafabadi and Ahmad Habibizad Navin</i>	
Application of Functional State Modelling Approach for Yeast <i>Sacharomyces cerevisiae</i> Fed-batch Fermentation Modelling.....	138
<i>Sandis Vilums, Emils Kozlinskis and Valters Brusbardis</i>	
Fetal ECG Representation Using Recurrence Plot Analysis.....	143
<i>Elif Tuba Celik, Bogdan Hurezeanu, Angela Digulescu and Madalina Mazilu</i>	
Data-Intensive Computing with Map-Reduce and Hadoop.....	147
<i>Shamil Humbetov</i>	

Web Based System for the Bee Colony Remote Monitoring	155
<i>Aleksejs Zacepins and Toms Karasha</i>	
High Speed Digital Filter Design using Register Minimization Retiming and Parallel Prefix Adders	159
<i>Deepa Yagain and Vijaya Krishna A.</i>	
Cloud Security Tactics: Virtualization and the VMM	165
<i>Panagiotis Kalagiakos and Margarita Bora</i>	
Designing an Active Band-Pass Filter with Tunable Transversal Element at 4-GHz using 0.2 m GaAs Technology	171
<i>Saeed Soleimany, Mohammad Reza Salehifar and Hassan Karbalaee</i>	
Determination of QoS Metrics in Wireless Sensor Networks by Using Queuing Theory	176
<i>Anar Rustamov</i>	
Coded OFDM Wireless Systems with Generalized Prefix	181
<i>Hakan Doğan, Hakan Yıldız, Todor Cooklev and Yusuf Acar</i>	
Evolution Mobile Wireless Communication and LTE Networks	185
<i>Tinatin Mshvidobadze</i>	
Conceptual Discrete Wavelet Transformation Speech Hashing for Content Authentication	192
<i>Mahdi Nouri, Nooshin Farhangian, Zahra Zeinolabedini and Nasim Fekr</i>	
Cloud Service for Comprehensive Project Management Software	198
<i>Ahmad Khan, Gundeep Singh Bindra, Rohan Arora, Nishant Raj, Darshan Jain and Dhruvad Shrivastava</i>	
Research on Forecasting Call Center Traffic through PCA and BP Artificial Neural Network	203
<i>Tao Liu and Lieli Liu</i>	
SoR based Request Routing for Future CDN	207
<i>Janaka Wijekoon, Erwin Harahap and Hiroaki Nishi</i>	
Cognitive Radio for Adaptive Modulation and Coding	212
<i>Sami H. O. Salih, Abbas Mohammed and Mamoun Suliman</i>	
SaPM: Switch-aware Process Mapping Model for Parallel Computing	217
<i>Yufei Lin, Yuhua Tang and Xinhai Xu</i>	
New Procedure for Discrimination of Model Parameter and Noise Variance Changes	221
<i>Theodor D. Popescu</i>	
With attackers wearing many hats, Prevent your “Identity Theft”	226
<i>Gundeep Singh Bindra, Dhruvad Shrivastava and Richa Seth</i>	

Fully Distributed Certificate Authority based on Polynomial over Elliptic Curve for MANET	231
<i>Ahmad Alomari</i>	
Cut-off Time Calculation for User Session Identification by Reference Length.....	236
<i>Jozef Kapusta, Michal Munk and Martin Drlik</i>	
Secure Data Aggregation in Wireless Multimedia Sensor Networks via Watermarking	242
<i>Ersin Elbasi and Suat Özdemir</i>	
Embedded Solution for Road Condition Monitoring Using Vehicular Sensor Networks.....	248
<i>Artis Mednis, Atis Elsts and Leo Selavo</i>	
Microstrip Patch Antenna Based on Photonic Crystal Substrate with Heterostructures at Terahertz Frequency	253
<i>Farhad alizadeh, Alireza maleki javan and Manoochehr kamyab hesari</i>	
Designing Yagi-Uda Antenna Fed by Microstrip Line and Simulated by HFSS	258
<i>Hassan Karbalaee, Mohammad Reza Salehifar and Saeed Soleimany</i>	
Multipurpose Smart SIM Card Based on Mobile Database and Location Dependent Query	263
<i>Hamid-Reza Firoozy-Najafabadi and Mohammad-Reza Feizi-Derakhshi</i>	
Estimation Models of Competition and Complementarity within Communication Technologies	268
<i>Nurilla Mahamatov and Suk Won Cha</i>	
Power-Aware Topology Generation for Application Specific NoC Design	273
<i>Suleyman Tosun, Yilmaz Ar and Suat Ozdemir</i>	
Wireless Platform for Multi-Channel RTD Measurements	279
<i>Amer Atta Yaseen</i>	
Weight-Based Fair Rate Allocation in Resilient Packet Ring.....	285
<i>Elyas Mohamadzadeh Kosari and Mohammad Hossein Yaghmaee Moghaddam</i>	
Data Processing in FPGA-based Systems	291
<i>Valery Sklyarov and Iouliia Skliarova</i>	
Genetic Algorithm Approach for the Prediction of Business Risks' Dynamics of Enterprise.....	299
<i>Gia Sirbiladze and Mikheil Kapanadze</i>	
An Empirical Study of Tracking Strategies of E-commerce Websites	304
<i>Wasin Treesinthuros</i>	
Automation of Business-Processes of an Election System	308
<i>Gia Surguladze, Nino Topuria, Ekaterine Turkia and George Basiladze</i>	
On the new multistage Fuzzy Technology to Investment Decisions	313
<i>Gia Sirbiladze, Irina Khutsishvili and Bezhan Ghvaberidze</i>	

Maximization of Firm’s profit Taking into Account Quality Criteria and Shelf Life of a Product	318
<i>Shorena Okujava</i>	
An Approach to Solving Some Management Problems Under Uncertainty	321
<i>Teimuraz Tsabadze and Tengiz Tsamalashvili</i>	
The model for Predicting the Competitiveness of Science Engineering Products	325
<i>Maslov A.V.</i>	
Social CRM: a New Solution for Relationship with Bank's Customers	328
<i>Mehrpooya Ahmadali Nejad and Seyyed Mohsen Hashemi</i>	
Problems and Prospects of Electronic Shops Development in Georgia	333
<i>David Zautashvili and Akaki Girgvliani</i>	
Information Access in the Globalised World	336
<i>Mariam Paposhvili and Aleksandra Suladze</i>	
Cloud computing for business	340
<i>Khayyam H. Masiyev, Ilkin Qasymov, Vusale Bakhishova and Mammad Bahri</i>	
Genetic Algorithm Approach in the Minimization of the Risk of Financial Portfolio	344
<i>P. Dvalishvili and B. Midodashvili</i>	
Intellectual Support System of EIA (Environmental Impact Assessment) Procedure in Region of Caspian Sea	347
<i>R.A. Karayev, K.A. Aliyev, M.A. Nagiev and N.E. Kazimova</i>	
Open-Access Journals (Free Valuable Information on the Web)	353
<i>Aref Riahi and Samaneh Khakmardan</i>	
A Key Component Extraction Method Based on HMM and Dependency Parsing	358
<i>Jianchu Kang, Songsong Pang, Jian Dong, Bowen Du and Jian Huang</i>	
The Role of Teacher in the Multimedia-based Foreign Language Classes	364
<i>Ketevan Gochitashvili</i>	
The Educational Communities’ social network	367
<i>Khayyam H. Masiyev, Nargiz Bayramova and Elvira Siraczade</i>	
Cognitive Methodology to Develop a Recovery Strategy for the Sturgeon Stocks of the Caspian Sea	371
<i>R.A. Karayev, A.I.L. Payne, N.Y. Sadikhova, K.A. Aliyev and A.N. Gasimli</i>	
Support Vector Domain Description for non-stationary data	376
<i>Foued Theljani, Kaouther Laabidi, Salah Zidi and Moufida Ksouri</i>	

The Usage of Malay Technological Terminologies in Malaysian Youth Institutions for Skills: Interests and Challenges	381
<i>Adenan Ayob</i>	
Analytical Solution of an Electrokinetic Flow in a Nano-Channel with Variable Physical Properties	385
<i>Mehdi Mostofi</i>	
Use of Informational Technologies in Study of the Spatial Structure of Glyprolines	389
<i>L.I.Ismailova, R.M.Abbasli, S.R.Akhmedova and N.A.Akhmedov</i>	
ZIPPER: The Holistic Spell Checker	393
<i>Lina Alhusaini</i>	
Analysis of the Relation between Turkish Twitter Messages and Stock Market Index	398
<i>Mehmet Ulvi Simsek and Suat Özdemir</i>	
Computer Determination of Preferred Conformations of Human Hemokinin-1	402
<i>U.T. Agaeva, G.A. Agaeva and N.M. Godjaev</i>	
Interactive Teaching Methods of Mathematical Physics by the methods of Computer Visualization	405
<i>S.T.Huseynov, N.V.Ibadov and A.A.Aslanov</i>	
A Web Application Tamper Proof Method Based on Text and Image Watermarking	410
<i>Zetao Jiang and Hongwu Zhang</i>	
Interactive Systems For Sign Language Learning	414
<i>Iurii Krak, Iurii Kryvonos and Waldemar Wojcik</i>	
Quantum Concepts in Information Retrieval	417
<i>M.Archuadze, G.Besiashvili, M.Khachidze and P.Kervalishvili</i>	
Decision Support System For Crisis Management Using Temporal Fuzzy Logic	421
<i>Ammar Alnahhas and Bassel Alkhatib</i>	
Recognition of the Patterns Represented as the Field Structures	426
<i>A.O. Chechel</i>	
Analyzing Reflector Effect on SAR Imaging System with a Proposed Functional Model	429
<i>Mojtaba Behzad Fallahpour and Hamid Deghani</i>	
Educational and Pedagogical Design of a Software Tool for Learning Postero-Anterior Cephalometric landmarking	433
<i>Francesco Maiorana and Rosalia Leonardi</i>	
Investigation of the Spatial Structure of Myomodulin E Molecule by Computer Modeling	438
<i>N.A. Akhmedov, L.N. Agaeva, R.M. Abbasli, L.I. Ismailova and N.M. Godjajev</i>	

Multi-objective Evolutionary Algorithm Based on Decomposition for Efficient Coverage Control in Mobile Sensor Networks	441
<i>Bara'a Ali Attea, Feyza Yıldırım Okay, Suat Özdemir and M. Ali Akcayol</i>	
Implementation of eLearning in Azerbaijan	447
<i>Leyla Muradkhanli</i>	
Modelling, simulation and monitoring the use of LabVIEW	450
<i>Štefan Koprda, Milan Turčáni and Zoltán Balogh</i>	
Underwater Scene Characterization Using Wavelet Packet Denoising and Adaptive Contrast Stretching	455
<i>Zhengmao Ye, Habib Mohamadian and Yongmao Ye</i>	
Computational Study of the Conformational Flexibility of the Amphibian Tachykinin Neuropeptides	461
<i>G.A. Agaeva</i>	
Application of Computer Technologies in Investigation of Spatial Structure of Peptide Molecules	465
<i>G.A. Akverdieva</i>	
Imitation Modeling of Competitive Market Equilibrium	470
<i>Khatuna Bardavelidze and Avtandil Bardavelidze</i>	
Methods of Evaluating Students in Different Countries	474
<i>Mehdi H.Soltani and Aloysat Q.Aliyev</i>	
Examining Characteristics, Obstacles, Reasons and Process of Smart Schools' Establishment	477
<i>Mehdi H.Soltani and Aloysat Q.Aliyev</i>	
Factor Analysis of Factors Affecting E-Learning Success from the Viewpoint of Virtual Students (Case Study of Islamic Republic of Iran)	481
<i>Aref Riahi, Hasan Khosravi and Samaneh Khakmardan</i>	
I Did Not Know Its Prohibited - Academic Dishonesty in Online Courses	485
<i>Yovav Eshet, Yehuda Peled and Keren Grinautski</i>	
Solving the Quadratic Assignment Problem with the Modified Hybrid PSO Algorithm	489
<i>Ali Safari Mamaghani and Mohammad Reza Meybodi</i>	
The Usage of Malay Technological Terminologies in Malaysian Youth Institutions for Skills: Interests and Challenges	495
<i>Adenan Ayob</i>	
Computing Infrastructure and Services Deployment for Research Community of Moldova	499
<i>P. Bogatencov, G. Secrieru and N. Iliuha</i>	

On the Method of Sustainable use of Soils by Regulation of Purchase Orders for Agricultural products	504
<i>Eyubova Svetlana, Pashayev Adalat and Sabziyev Elkhan</i>	
Comparison of the Efficiency of Principal Component Analysis and Multiple Linear Regression to Determine Students' Academic Achievement.....	507
<i>Mehtap Erguven</i>	
Computer Modeling of Hemokinins using Molecular Dynamics Method: Hemokinin 1 (human).....	512
<i>N.M. Qodjayev, B.M. Qasimov and U.T. Agayeva</i>	
Using of Information and Communication Technologies in Learning and Teaching of Physics in Universities: Few Parametric Model for the Dynamics of Vibrations of Diatomic Molecules	515
<i>R. Qadmaliyev, B.M. Qasimov and N.M. Qodjayev</i>	
Teacher's Role in the Process of Call in Language Teaching	519
<i>Ali Shahintash and Mehmet Shahiner</i>	
Author Index	Follows page 521

SESSION 1

Information Science and Application

AICT 2012

Ontology Creation for an Educational Center

Selma DILEK, Hacer KARACAN, Nesa JAHANGIRI, Shima AFZALI

Gazi University
Faculty of Engineering, Dept. of Computer Engineering
Ankara, Turkey

Abstract—The introduction of the Semantic Web in 2001 led the way to many new ideas and technologies in computing. Data handling and sharing gained a semantic level rather than structural, facilitating the development of innovative and more helpful end-user applications. Semantics of Web information is formally defined in ontologies, making it possible for machines to interpret data content more effectively. In this study, we aimed to apply this new technology and create a useful application for an educational center. This paper presents our work, issues we encountered, our approach to resolving those issues, as well as our findings and results.

Keywords; Semantic Web, ontology, OWL, Jena.

I. INTRODUCTION

After Berners-Lee, Hendler, and Lassila introduced the concept of the Semantic Web in 2001 [1], it led the way to many new ideas and technologies in computing. Data handling and sharing gained a semantic level rather than structural, facilitating the development of innovative and more helpful end-user applications [2].

According to W3C Semantic Web Activity, the Semantic Web “is an evolving extension of the World Wide Web in which the semantics, or meaning, of information on the Web” is formally defined. Formal definitions are captured in ontologies, making it possible for machines to interpret and relate data content more effectively. The principal technologies of the Semantic Web include the Resource Description Framework (RDF) data representation model and the ontology representation languages RDF Schema and Web Ontology Language (OWL)” [3].

Ontologies can be used for encoding domain knowledge and providing a semantic representation of the information in databases. Moreover, data integration is facilitated by introduction of semantics. Ontologies also help resolve discrepancy between data and user queries, because they enable performing semantic queries [4]. This property is of special interest to our project, in which we aimed to apply this new technology and create a useful application for one of the most important educational centers in Ankara, Turkey. We anticipated that there would be a need for various complex semantic queries over our database, and defining data in ontology provides this possibility.

This paper presents our work in the aforementioned project, issues we encountered, our approach to resolving those issues, as well as our findings and results.

II. PROBLEM DESCRIPTION

The subject of our study was creating an ontology for the educational center KEÇMEK (Keçiören Belediyesi Sanat ve Meslek Eğitimi Kursları – Turkish for Keçiören Municipality Educational Courses in Arts and Professions). KEÇMEK is a professional educational institution spread over 14 centers, which offers a vast number of courses in 70 different fields in arts and vocation certified by the Turkish Ministry of Education. This important educational institution does not even have a web site, making it really hard to advertise its services and reach out to the larger number of citizens. The only marketing is done via posters that are displayed inside the municipality building. Also, there is no system for better organization of services and information, as well as communication between centers and departments. Such a huge institution needs a better online system for presentation of its services, student applications, information gathering, interdepartmental communication, etc.

III. CREATION OF ONTOLOGY – WORK PROGRESS

Our ontology development tool of choice is Protégé - a free, open source ontology editor and knowledge-base framework from Stanford Center for Biomedical Informatics Research. Our language of choice is OWL-DL, because it corresponds to a well-studied description logic, and permits efficient reasoning support.

Our work progress was the following:

1. We defined all components and relations in our ontology (See figure 1)

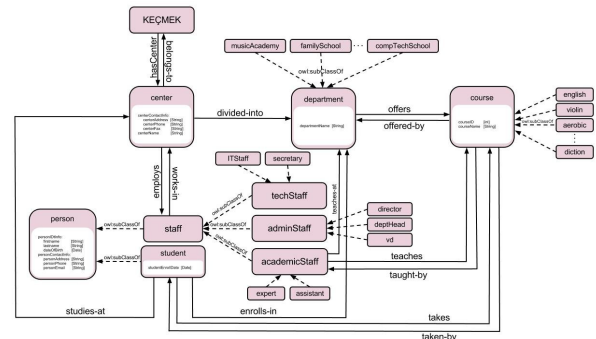


Figure 1. Components and Relations

2. We defined classes,

3. We arranged the classes in a subclass-superclass hierarchy, and defined the disjoint classes,

For instance, this is how the *screenwriting* class (a subclass of the class *course*) is created:

```
<owl:Class rdf:ID="screenwriting">
  <rdfs:subClassOf>
    <owl:Class rdf:ID="course"/>
  </rdfs:subClassOf>
  <owl:disjointWith>
    <owl:Class rdf:ID="familyEducation"/>
  </owl:disjointWith>
  <owl:disjointWith>
    <owl:Class rdf:ID="english"/>
  </owl:disjointWith>
  <owl:disjointWith>
    <owl:Class rdf:ID="violin"/>
  </owl:disjointWith>
  .
  .
```

} Defining its superclass.

} Defining disjoint classes. Here, we give only a few examples for the sake of brevity.

4. We defined Object Properties, set restrictions, domains and ranges for these properties as necessary, and defined inverses where applicable,

As an example, we give a part of the code where we defined Object Property *takes* (from: *student takes course*), and specified its inverse (*taken-by*), its domain (*student*) and range (*course*):

```
<owl:ObjectProperty rdf:about="#takes">
  <owl:inverseOf>
    <owl:ObjectProperty rdf:about="#taken-by"/>
  </owl:inverseOf>
  <rdfs:domain rdf:resource="#student"/>
  <rdfs:range rdf:resource="#course"/>
</owl:ObjectProperty>
```

As an example how we used owl:Restriction, we take a look at the rest of the code for *screenwriting* class:

```
<owl:Class rdf:ID="screenwriting">
  <rdfs:subClassOf>
    <owl:Restriction>
      <owl:onProperty>
        <owl:ObjectProperty rdf:ID="offered-by"/>
      </owl:onProperty>
      <owl:someValuesFrom>
        <owl:Class rdf:ID="cinemaWorkshop"/>
      </owl:someValuesFrom>
    </owl:Restriction>
  </rdfs:subClassOf>
</owl:Class>
```

Here, we use owl:Restriction to define that *screenwriting* course is offered strictly by *Cinema Workshop* department by putting a restriction on Object Property *offered-by* that ties a *course* to a *department*.

5. We defined Datatype Properties where needed. For instance, this is how we defined Datatype Property *centerContactInfo*:

```
<owl:DatatypeProperty rdf:ID="centerContactInfo">
  <rdfs:domain rdf:resource="#center"/>
</owl:DatatypeProperty>
```

And this is how we defined Datatype Property *centerAddress* which is a sub-property of *centerContactInfo*, and since there can be only one address, we define it as a Functional Property:

```
<owl:FunctionalProperty rdf:ID="centerAddress">
  <rdfs:subPropertyOf rdf:resource="#centerContactInfo"/>
  <rdf:type
    rdf:resource="http://www.w3.org/2002/07/owl#DatatypeProperty"/>
  <rdfs:range
    rdf:resource="http://www.w3.org/2001/XMLSchema#string"/>
</owl:FunctionalProperty>
```

6. We defined cardinality restrictions, such as:

- 6.1. A center must have at least one department,
- 6.2. A department must offer at least one course,
- 6.3. A course must be attended by at least three students,
- 6.4. Each member of academic staff must teach at least one course, etc.

As an example of setting cardinality restrictions, we give the restriction we defined on the *student* class stating that a student must take at least one course (the Object Property *takes* is defined to have the *student* class as its domain, and the *course* class as its range):

```
<owl:Class rdf:ID="student">
  <rdfs:subClassOf>
    <owl:Restriction>
      <owl:onProperty rdf:resource="#takes"/>
      <owl:minCardinality
        rdf:datatype="xsd:nonNegativeInteger"> 1
      </owl:minCardinality>
    </owl:Restriction>
  </rdfs:subClassOf>
</owl:Class>
```

7. We inserted instances of our classes (centers, departments, courses, students, staff, etc.).

As an example of creating instances, we give an instance of the class *student*. For the sake of brevity, we leave out the definitions of some of the student class's Datatype Properties such as: address, phone, occupation, etc.:

```
<student rdf:ID="student_5">
  <firstname
    rdf:datatype="http://www.w3.org/2001/XMLSchema#string">
```



```

Sema
</firstname>
<lastname
rdf:datatype="http://www.w3.org/2001/XMLSchema#string">
  Alpaslan
</lastname>
<dateOfBirth
rdf:datatype="http://www.w3.org/2001/XMLSchema#date">
  1994-08-10
</dateOfBirth>
...
</student>

```

8. After we finished creating the ontology, we performed consistency check over the ontology using Pellet 1.5.2 reasoner, and confirmed that all ontology components are consistent.

IV. QUERYING THE ONTOLOGY

Our goal was to create an application that will utilize our ontology:

- For prospective students: presentation and information system that enables browsing through existing centers, departments and offered courses, and application form.
- For administration: easy access to employee and student records and their contact information.

We uploaded our ontology as .owl file to <http://kecmek.host22.com> (See Figure 2). The direct link is <http://kecmek.host22.com/kecmek.owl>. In order to perform queries over the kecmek.owl, we used Apache Jena™ (a Java framework for handling Semantic Web applications) within NetBeans IDE 7.1.2. Using Jena libraries, we successfully read the ontology file from the web into an ontology model, which we used to access the components of our ontology and perform various queries.



Figure 2. the domain we registered for our study

A. Declaring Namespace:

```

public static String BASE =
    "http://kecmek.host22.com/kecmek.owl";
public static String NS = BASE + "#";

```

B. Reading the ontology into an ontology model m:

```

OntModel m = getKecmekOntology();
protected OntModel getKecmekOntology()
{
    OntModel m =
    ModelFactory.createOntologyModel(OntModelSpec.OWL_DL
    _MEM); m.read(BASE);
    return m;
}

```

And we also give an example query on our model m. Here, users are asked to input what they are looking for (e.g. center, department, etc.), and the result of our query displays the elements of the searched item:

```

System.out.print("What are you looking for?: ");
String input = EasyIn.readString();
OntClass query = m.getOntClass(NS + input);
System.out.println("Search results:");
for (Iterator instances = query.listSubClasses();
instances.hasNext(); )
{
    OntClass temp =
    m.getOntClass(instances.next().toString());
    System.out.println(temp.getComment(null));
}

```

V. RESULTS AND DISCUSSION

As we mentioned in the introduction, ontologies facilitate overcoming inconsistencies between data and user queries, because they enable performing semantic queries. An example of a semantic query that could be performed over our ontology is “finding an English language course that is offered by a specific department and taught by a specific lecturer”. Therefore, choosing to build ontology over building a standard database for our studies, proved to be a wise decision.

Furthermore, we chose OWL over RDF Schema, because with RDF Schema we can only represent concepts and simple relations between them, whereas OWL adds more semantics to the schema, allowing us to state far more complex relations and properties.

Protégé proved to be a very user-friendly and helpful tool for creating ontologies. Creating classes, properties, instances and adding restrictions is straightforward and it offers some helpful shortcuts such as stating that all sibling classes are disjoint. Furthermore, Protégé offers a built-in OWL reasoner, so we were able to easily perform the consistency check of our ontology. Jena proved to be a powerful framework for handling data processing in Semantic Web applications. As for which IDE to use in order to build an application over the ontology, we can recommend NetBeans IDE which proved to be quite adequate and user-friendly, since we encountered certain problems when we tried to use eclipse. We were able to successfully access the components of our ontology and perform various queries (See Fig. 3).

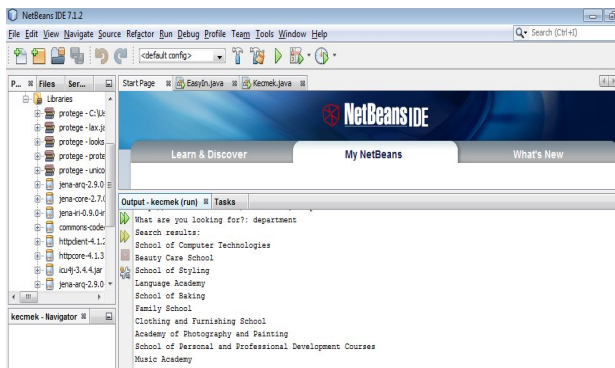


Figure 3. Some of our Queries

During the creation of ontology, we had also come across another problem, regarding attaching specific courses to their departments such that, for instance, English course can be offered only by Language Academy department. We solved this problem by putting a restriction on Object Property *offered-by* that ties a *course* to a *department*. For instance, we defined that the English course is offered strictly by Language Academy department by writing:

```
<owl:Class rdf:ID="english">
  <rdfs:subClassOf rdf:resource="#course"/>
  <rdfs:subClassOf>
    <owl:Restriction>
      <owl:onProperty rdf:resource="#offered-by"/>
      <owl:hasValue
        rdf:resource="#languageAcademy"/>
    </owl:Restriction>
  </rdfs:subClassOf>
</owl:Class>
```

As a result, we can conclude that Semantic Web technology offers new possibilities that will change the way we organize and process information. Adding semantics to information facilitate execution of much more meaningful and complex queries. Ontologies make it possible for machines to interpret and relate data content more effectively, thus, enabling the development of more useful end-user applications.

REFERENCES

- [1] T. Berners-Lee, J. Hendler, and O. Lassila, "The Semantic Web", Scientific American, 2001,
- [2] A. Batzios, P. A. Mitkas, "WebOWL: A Semantic Web search engine development experiment," Expert Systems with Applications, 39: 5052–5060, 2012
- [3] A. Sheth, C. Henson, and S. Sahoo, "Semantic Sensor Web," IEEE Internet Computing, 78–83, 2008.
- [4] L. Ma, J. Mei, Y. Pan, K. Kulkarni, and A. Fokoue, "Semantic Web Technologies and Data Management," W3C papers, <http://www.w3.org/2007/03/RdfRDB/papers/ma.pdf>, 2007, last accessed on 14/07/2012
- [5] G. Antoniou and F. van Harmelen, „A Semantic Web Primer,“ The MIT Press, Cambridge, Massachusetts, London, England, 2004.
- [6] H. Karacan, Semantic Web - Lecture Notes, <http://ceng.gazi.edu.tr/~hkaracan/bm521.html>, last accessed on 16/05/2012
- [7] <http://protege.stanford.edu/>, last accessed on 16/05/2012

- [8] <http://jena.apache.org/>, last accessed 16/05/2012
- [9] <http://protegewiki.stanford.edu/wiki/BuildingSemanticWebApplications>, last accessed on 16/05/2012
- [10] <http://netbeans.org/>, last accessed on 16/05/2012
- [11] <http://stackoverflow.com/>, last accessed on 16/05/2012

New novel idea for Cloud Computing: How can we use Kalman filter in security of Cloud Computing

Mehdi Darbandi¹, Pariya Shahbazi^{2}, Saeed Setayesh³, Ole-Christoffer Granmo⁴*

¹Iran University of Science and Technology (IUST); Tehran, Iran

^{2,4}University of Agder (UiA); Grimstad, Norway; (corresponding author: Paria Shahbazi)*

³International Branch of Ferdowsi University of Mashhad; Mashhad, Iran

Abstract— Cloud is a virtual image about some amount of undefined powers, that is widespread and had unknown power and inexact amount of hardware and software configurations, and because of we have not any information about clouds location and time dimensions and also the amounts of its sources we tell that Cloud Computing. This technology presents lots of abilities and opportunities such as processing power, storage and accessing it from everywhere, supporting, working – team group - with the latest versions of software and etc., by the means of internet. On the other hand, in such a large scale networks we should consider the reliability and powerfulness of such networks in facing with events such as high amount of users that may login to their profiles simultaneously, or for example if we have the ability to predict about what times that we would have the most crowd in network, or even users prefer to use which part of the Cloud Computing more than other parts – which software or hardware configuration. With knowing such information, we can avoid accidental crashing or hanging of the network that may be cause by logging of too much users. In this paper we propose Kalman Filter that can be used for estimating the amounts of users and software's that run on cloud computing or other similar platforms at a certain time. After introducing this filter, at the end of paper, we talk about some potentials of this filter in cloud computing platform. In this paper we demonstrate about how we can use Kalman filter in estimating and predicting of our target, by the means of several examples on Kalman filter.

Keywords- Cloud computing and its influences, Security, Kalman Filter, Estimation and prediction.

I. INTRODUCTION

History of computer science has had fundamental changes. For example in first generation of computers (from 1945 until 1956), vacuum tube was used in computers. In 2nd generation (from 1956 until 1963) by the invention of transistors, these tiny devices were used in computers. After that, integrated circuits were used in 3rd generation of computers (1963 to 1971). Finally in 4th generation (since 1971 up to now), along with technology advancement, Large Scale Integration (LSI) circuits, Very Large Scale Integration (VLSI) and Ultra Large Scale Integration (ULSI) were used in computers. Nowadays new technology that is named as Cloud Computing is creating a new era in computer industry and processing power. By reviewing the historical points, we can understand that the idea of Cloud Computing taken from this fact: When the current user or users don't require the processing resources, these resources can be assigned to other users. Most simple definition of Cloud Computing is: Access to enormous resources and processing powers even, through cheap computers.

One of the major benefits of this technology is sharing resources – software and hardware resources - and the

ability to simultaneously working on specific projects or files. For example you want to do some processing on DNA of person(s). If you are user of cloud computing, at first you can use processing power of cloud computing that is more reliable and efficient, because your system may not have enough processing power. Secondly, you can do the above operations with many experts – sharing your task - simultaneously and find instant results of each other. Also, before this technology revealed, you were restricted to use only processing power of your own system. Also, regularly you must update the hardware and software of your system, in order to be able to access the latest software version and also have required hardware to support this software, that's very costly. Also if you want to install this software on more computers; you need to purchase more licenses of that software.

We don't have such problems in Cloud Computing, because every time you log out from your account and logging to it next time, you'll see the latest update of the software, without need to developing the hardware of your system, because the required hardware is provided by cloud resources. Also you don't pay additional costs for this software developing – buying licenses or pay for updating the software.

In ordinary systems, if you have high sensitive and important information on your computers, you must always update antivirus and firewall of your own system, to avoid from data loss or manipulation. If you forget doing this action (updating), it is possible to accidentally your system be attacked and your information became destroyed or manipulated. But in Cloud Computing, when you login to your account, you will see security system is up-to-date with the latest version.

Another benefit of cloud computing is, supporting (recognizing) all different formats of the files. For example, in ordinary systems, you may download a file – from your e-mail or from a website - that has unknown format for your machine and you don't know which software can open it; but in Cloud Computing this file will be open with associated and applicant software –automatic recognition of that file type and devoting it the exact software for running is occurred in cloud computing.

Moreover, for instance if you have a company, you can transfer internal network of your company –your server database - on Cloud Computing to enjoy more speed and processing power, and also if you use server, you will economize in budget and only pay power consumption and maintenance costs.

These are just part of the great performance of new technology, known as Cloud computing that is named also as "the next big thing" [1-9].

With the rapid development of Internet and Cloud computing, there are more and more network resources. Sharing, management and on-demand allocation of network resources are particularly important in Cloud computing. Platform as a Service (PaaS) is one of the key services in Cloud computing. PaaS is very attractive for schools, research institutions and enterprises which need reducing IT costs, improving computing platform sharing and meeting license constraints. However, nearly all current available cloud computing platforms are either proprietary or their software infrastructure is invisible to the research community except for a few open-source platforms. For universities and research institutes, more open and testable experimental platforms are needed in a lab-level with PCs.

Cloud Computing is a promising paradigm designed to harness the power of networks of computers and communications in a more cost effective way. Clouds provide elastic capacity to serve a wide and constantly expanding range of information processing needs, including government, military, business and education. The Cloud Computing paradigm is maturing rapidly and is being considered for adoption in government and business platforms. Open source systems refer to software systems whose source code is available, allowing for immediate incorporation of improvements and adaptations of the system by its users.

II. CONSIDERING HIGH IMPACTS OF CLOUD COMPUTING ON DIFFERENT INDUSTRIES

In two past sections of the paper, we define some of the basic and fundamental principles of cloud and also we tell about some of its advantageous. Now we want imply into, the major applications of this technology. After that when we understand the importance of this technology, we introduce Kalman Filter; which can be used for prediction and estimation of different parameters in cloud platform.

The main reason for using cloud computing is, humans in the current era wants to use wireless and high speed communications [6]. The first commercially provider of cloud computing was Amazon. Another reason for this amount of usage is that, we prefer not to know about the volume of processors, hardware and servers and only wants to face with very gigantic ability in storage and processing capability [9]. The third reason for using of cloud computing instead of traditional systems is that, these days people want to use social sites, and these sites should have enough capability to support the users; we shouldn't have any crash or low speed when we use them. For example Facebook has more than 850 million members these days and the Facebook Team decides recently that it's better to migrate to Cloud Computing platform, because it had more storage and processing power [14]. Another reason is that user(s) can decide with whom and with which places in all over the world, they wants to share his/her information and files; we don't have such abilities in traditional systems, also in traditional systems we have the restriction on the volume of data that's stored [11]. Another reason is that, a programmer doesn't need to program the servers and computers for particular users or organizations, but every people can use cloud computing with his/her special uses that wants. Furthermore, it doesn't need any specialty in using, it doesn't have any software and hardware requirement, and in

cloud computing all kind of files were recognizable – automatic format recognition [9].

Google Company wants to publish Chrome operating system, and with that it wants to visit their users in cloud. They believe that by employing that technology they can better satisfy their users. Furthermore, Google recently has published the new version of Google Docs, which is very powerful software package for cloud users. Also Microsoft Company decides to introduce Microsoft Azure – that's an OS which is based on cloud computing, because they fear about losing their users. Also General Service Administration (GSA); because of having so many visitors that visit their sites every day, they fearing about crashing or hanging of it, as a result they decide to migrate to cloud computing. National Aeronautics and Space Administration (NASA) builds NEBULA, which is cloud computing platform and with using of that platform, people can participate in NASA missions and tell their ideas and even their suggestions; moreover NASA search for better storage and processing ability from this technology. The NASA Headquarter has publishes on their recent report, that they decide to implement International Space Station platform on Cloud Computing. Also Department of Interior, which is provider of lots of services for organizations decide to present some services by cloud computing, because they think it's more convenient. Department of Health and Human Services wants to employ new platform which is based on cloud computing to be able to give better, faster and more efficient services. Census Bur, which is the provider and supporter of SaaS services in Salesforce site - that is giving services to millions of people every day - decide to present new products based on Cloud platform. Also The White House decide to migrate to cloud computing technology because they think in that way they can do their tasks such as e-voting, having conversation with their citizens, and their internal networks; more easily and faster [1, 5,15].

Also United Kingdom Governments run G-Cloud networks for themselves, in order to have more precise in their works. Europeans use cloud computing in public and private sections such as: Management of public sector housing, transportation service networks, Census, Economic development, Health services, and Contracting and education services [1].

Also in Denmark with two pilot projects, Digitalise'r.dk and NemHandel, they evaluate cloud computing for their users and after that when they became satisfy about this technology they decide to establish new platform for their governments based on cloud computing. One of the pioneers in this technology are Japanese with their cloud, that known as Kasumigaseki Cloud and built in government-industry area of Tokyo, they wants to improve cloud for their public uses, also they named this technology and area that implement it as a pilot project as “green environment”. In China, especially in north of this country the government with the project that named as: “The Yellow River Delta Cloud Computing Center”, wants to establish governmental based services on this technology. Even in Wuxi city, the government established a company for manufacturing resources based on cloud computing technology. In Thailand, The Government Information Technology Services (GITS), design and construct private cloud for

public sectors; and they want to, as soon as possible present so many services for civilians and private organizations by the means of cloud computing. In Vietnam, IBM Company with their government and universities wants to develop the new laws, for presenting cloud for public and private sections. In New Zealand, they search for the better and more efficient uses and potentials of cloud computing [1, 7]. Now we want to consider the impact of this technology on minor companies and organizations. YouTube Company, which is in 2006, has a daily increasing about 30 million in their web-pages, nowadays decide to use cloud computing technology for better and faster searching and hosting, because as they told in their report we have more intelligent algorithms in cloud platform. Although if such company doesn't use cloud computing, they should pay lots of money for maintaining and operating of their servers, also lots of money for upgrading their software and hardware configurations [13].



Fig. 1: Every People Can Share Their Tasks and Projects with Others, They Can Get Helps and New Ideas.

SmugMug Company which is a site for sharing images; for better and faster services and also because of they doesn't have enough storage space decide to migrate to cloud computing. Or as an example Google Company prepares services such as Google App Engine which promise you to run your applications on Google cloud resources. As such, you can share your files and processes with anyone you want or even you can share your files with all the peoples around the world. Also Google and IBM decide to construct and run new networks for universities based on cloud computing; so that universities can do their researches with help of each other simultaneously, even faster and with more precise; because the students and professors from other universities can take part in other research topics and tell their ideas. The universities which are used this technology are: University of Washington, Carnegie-Mellon University, MIT, Stanford University, The University of California at Berkeley, Maryland University [7].

The smart home is a concept of the pervasive computing, and it gradually becomes significant for the people living in the high technology area. For numerous data and complex control bring about a much heavy burden on the local computers, and it is difficult for the users to obtain the information of the smart home. In this paper, we propose the

smart home structure based on Cloud Computing, which helps to reduce local workload and the users obtain the real time information through Web browser directly.

Nasdaq Company, which is had lots of data about stock and funds, wants to share and sell their information; but they are anxious about using of servers and their storage capacity; thus they decide to use Amazon S3 service, which is cloud based platform. Recently, minor companies such as Nimbus and Eucalyptus present storage and processing powers by getting fees [1, 13].

Sun Micro Systems Company now involved in constructing new data centers for hosting applications and users of cloud computing. This company wants to establish many sites all over the world to hosting user's application and avoiding from failure. Also these days science networks such as My experiment and nanoHub migrates to cloud computing, because they want to had better and more efficient and convenient communications with their fans. Also cloud computing can present banking services, it means you take some services by paying their fees and after that you improve that service or add some features to that service and sell it to other users. Also, iPhone company build their new Mobile Phone based on Cloud computing platform, the providers of this product think that with using this technology they can implement chipper Microcontrollers on the board of this device, but instead they can gain more processing power, only because they use cloud computing technology. Finally, writers of this paper predict in near future lots of people involved in this technology. Engineers and designers must play their role more accurately to get the interest of more users and provide them their needs [1 to 5]. In the next section we discuss about basic concepts of Kalman Filter and introduce it briefly. We can use this algorithm for estimation and prediction the amount of users that logging into their profiles at certain time; furthermore

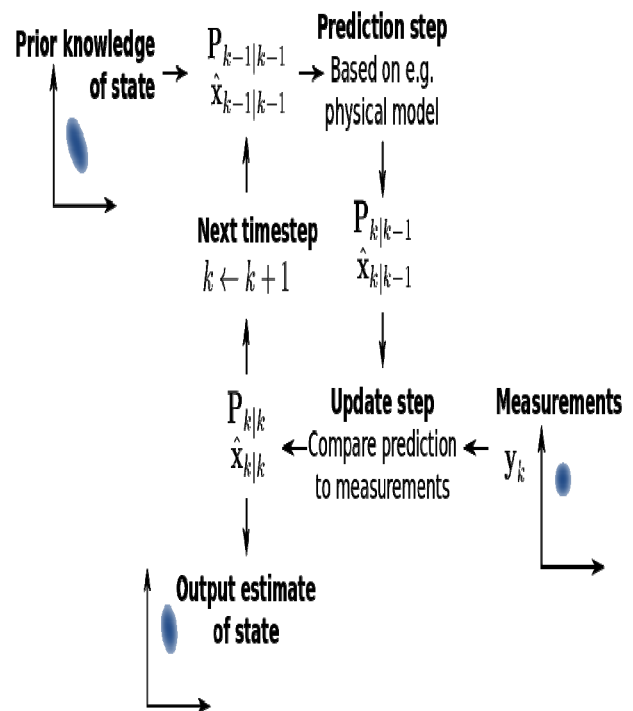


Fig. 2: The Kalman filter keeps track of the estimated state of the system and the variance or uncertainty of the estimate. The estimate is updated using a state transition model and measurements. $\hat{x}_{k|k-1}$ Denotes the estimate of the system's state at time step k before the k -th measurement y_k has been taken into account; $P_{k|k-1}$ is the corresponding uncertainty.

we can use it for security purposes.

III. AN INTRODUCTION INTO KALMAN FILTER

In 1960, R.E. Kalman published his famous paper describing a recursive solution to the discrete-data linear filtering problem. Since that time, due in large part to advances in digital computing, the Kalman filter has been the subject of extensive research and application, particularly in the area of autonomous or assisted navigation. The Kalman filter is a set of mathematical equations that provides an efficient computational (recursive) means to estimate the state of a process, in a way that minimizes the mean of the squared error. The filter is very powerful in several aspects: it supports estimations of past, present, and even future states, and it can do so even when the precise nature of the modeled system is unknown. The Kalman filter has numerous applications in technology.

A simple and ubiquitous example is the phase-locked loop, used in FM radios and most electronic communications equipment. Another common application is for guidance, navigation and control of vehicles, particularly aircraft and spacecraft.

The Kalman filter uses a system's dynamics model (i.e., physical laws of motion), known control inputs to that system, and measurements (such as from sensors) to form an estimate of the system's varying quantities (its state) that is better than the estimate obtained by using any one measurement alone. As such, it is a common sensor fusion algorithm.

All measurements and calculations based on models are estimates to some degree. Noisy sensor data, approximations in the equations that describe how a system changes, and external factors that are not accounted for introduce some uncertainty about the inferred values for a system's state. The Kalman filter averages a prediction of a system's state with a new measurement using a weighted average. The purpose of the weights is that values with better (i.e., smaller) estimated uncertainty is "trusted" more. The weights are calculated from the covariance, a measure of the estimated uncertainty of the prediction of the system's state. The result of the weighted average is a new state estimate that lies in between the predicted and measured state, and has a better estimated uncertainty than either alone. This process is repeated every time step, with the new estimate and its covariance informing the prediction used in the following iteration. This means that the Kalman filter works recursively and requires only the last "best guess" - not the entire history - of a system's state to calculate a new state.

When performing the actual calculations for the filter (as discussed below), the state estimate and covariance's are coded into matrices to handle the multiple dimensions involved in a single set of calculations. This allows for representation of linear relationships between different state variables (such as position, velocity, and acceleration) in any of the transition models or covariances.

The Kalman filter is used in sensor fusion and data fusion. Typically, real-time systems produce multiple sequential measurements rather than making a single measurement to obtain the state of the system. These multiple measurements

are then combined mathematically to generate the system's state at that time instant.

As an example application, consider the problem of determining the precise location of a truck. The truck can be equipped with a GPS unit that provides an estimate of the position within a few meters. The GPS estimate is likely to be noisy; readings 'jump around' rapidly, though always remaining within a few meters of the real position. The truck's position can also be estimated by integrating its speed and direction over time, determined by keeping track of wheel revolutions and the angle of the steering wheel. This is a technique known as dead reckoning. Typically, dead reckoning will provide a very smooth estimate of the truck's position, but it will drift over time as small errors accumulate. Additionally, the truck is expected to follow the laws of physics, so its position should be expected to change proportionally to its velocity.

In this example, the Kalman filter can be thought of as operating in two distinct phases: predict and update. In the prediction phase, the truck's old position will be modified

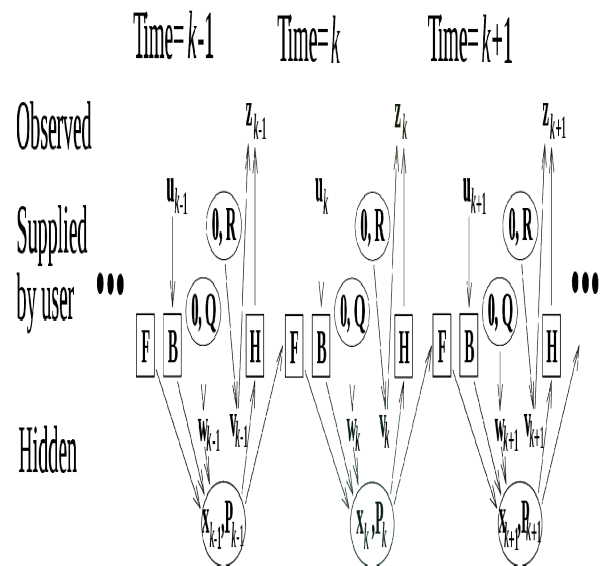


Fig. 3: Model underlying the Kalman filter. Squares represent matrices. Ellipses represent multivariate normal distributions (with the mean and covariance matrix enclosed). Unenclosed values are vectors. In the simple case, the various matrices are constant with time, and thus the subscripts are dropped, but the Kalman filter allows any of them to change each time step.

according to the physical laws of motion (the dynamic or "state transition" model) plus any changes produced by the accelerator pedal and steering wheel. Not only will a new position estimate be calculated, but a new covariance will be calculated as well. Perhaps the covariance is proportional to the speed of the truck because we are more uncertain about the accuracy of the dead reckoning estimate at high speeds but very certain about the position when moving slowly. Next, in the update phase, a measurement of the truck's position is taken from the GPS unit. Along with this measurement come some amount of uncertainty, and its covariance relative to that of the prediction from the previous phase determines how much the new measurement will affect the updated prediction. Ideally, if the dead reckoning

estimates tend to drift away from the real position, the GPS measurement should pull the position estimate back towards the real position but not disturb it to the point of becoming rapidly changing and noisy. As another example application consider the use of Kalman filter in computer vision, Data fusion using a Kalman filter can assist computers to track objects in videos with low latency (not to be confused with a low number of latent variables). The tracking of objects is a dynamic problem, using data from sensor and camera images that always suffer from noise. This can sometimes be reduced by using higher quality cameras and sensors but can never be eliminated, so it is often desirable to use a noise reduction method. The iterative predictor-corrector nature of the Kalman filter can be helpful, because at each time instance only one constraint on the state variable need be considered. This process is repeated, considering a different constraint at every time instance. All the measured data are accumulated over time and help in predicting the state. Video can also be pre-processed, perhaps using a segmentation technique, to reduce the computation and hence latency.

IV. CONCLUSION

In this article, we tell about basic definitions and concepts of cloud computing. Also we tell about uses of this technology in different societies and industries and also about how they use it and about their future plans. An important factor in using cloud computing and migrating into this network is its security and durability. In this paper the authors propose Kalman filtering for increasing security and durability of such networks for the first time. If we implement this algorithm on such networks – for example on the edge of such networks, we can estimate and predict the amount of users that use the resources – software and hardware resources – at anytime. Also we can estimate and predict the amount of users that logging onto a certain account, and by the means of that we can avoid surveillance entering of bad users – we estimate the location of user by the previous location and its background data. Also we can use it for estimating the amount of user that use a certain application on such networks, and by knowing that amount we can improve power of our network to be able to support our users. Furthermore by using this algorithm we can increase the security of this technology, by estimating and predicting the point of presence of bad users. In this paper we demonstrate about how can we use Kalman filter in estimating and predicting of our target, by the means of several examples on Kalman filter.

REFERENCES

[1] David C. Wyld; “the cloudy future of government IT: cloud computing and the public sector around the world”, *IJWesT*, Vol. 1, Num. 1, Jan. 2010.

[2] Jean-Daniel Cryans, Alain April, Alain Abran; “criteria to compare cloud computing with current database technology, R. Dumke et al. (Eds.): *IWSM / MetriKon / Mensura 2008*, LNCS 5338, pp. 114-126, 2008.

[3] Anil Madhavapeddy, Richard Mortier, Jon Crowcroft, Steven Hand; “multiscale not multicore: efficient heterogeneous cloud computing”, published by the British Informatics Society Ltd. *Proceedings of ACM-BCS Visions of Computer Science 2010*.

[4] Harold C. Lim, Shivnath Babu, Jeffrey S. Chase, Sujay S. Parekh; “automated control in cloud computing: challenges and opportunities”, *ACDC’09*, June 19, Barcelona, Spain.

[5] N. Sainath, S. Muralikrishna, P.V.S. Srinivas; “a framework of cloud computing in the real world”; *Advances in Computational Sciences and Technology*, ISSN 0973-6107, Vol. 3, Num. 2, (2010), pp. 175-190.

[6] Kyle Chard, Simon Caton, Omer Rana, Kris Bubendorfer; “social cloud: cloud computing in social networks”

[7] G. Bruce Berriman, Eva Deelman, Paul Groth, Gideon Juve; “the application of cloud computing to the creation of image mosaics and management of their provenance”;

[8] Roy Campbell, Indranil Gupta, Michael Heath, Steven Y. Ko, Michael Kozuch, Marcel Kunze, Thomas Kwan, Kevin Lai, Hing Yan Lee, Martha Lyons, Dejan Milojicic, David O’Hallaron, Yeng Chai Soh; “open cirrus™ cloud computing testbed: federated data centers for open source systems and services research”

[9] Rajkumar Buyya, Chee Shin Yeo, Srikumar Venugopal, James Broberg, Ivona Brandic; “cloud computing and Emerging IT platforms: Vision, Hype, and Reality for delivering computing as the 5th utility”

[10] Lamia Youseff, Maria Butrico, Dilma Da Silva; “toward a unified ontology of cloud computing”

[11] Daniel A. Menasce, Paul Ngo; “understanding cloud computing: experimentation and capacity planning”; *Proc. 2009, Computer Measurement Group Conf. Dallas, TX. Dec. 2009*.

[12] Won Kim; “cloud computing: today and tomorrow”; *JOT*, Vol. 8, No. 1, Jan-Feb 2009.

[13] Richard Chow, philippe Golle, Markus Jakobsson, Elaine Shi, Jessica Staddon, Ryusuke Masuoka, Jesus Molina; “controlling data in the cloud: outsourcing computation without outsourcing control”; *CCSW’09*, Nov. 13, 2009, Chicago, Illinois, USA.

[14] Bo Peng, Bin Cui, Xiaoming Li; “implementation issues of a cloud computing platform”; *Bulletin of the IEEE computer society technical committee on data engineering*.

[15] Daniel Nurmi, Rich Wolski, Chris Grzegorzczak, Graziano Obertelli, Sunil Soman, Lamia Youseff, Dmitrii Zagorodnov; “the eucalyptus open-source cloud computing system”



Particle Swarm Intelligence as a New Heuristic for the Optimization of Distributed Database Queries

Tansel Dokeroglu, Umut Tosun, Ahmet Cosar

Department of Computer Engineering Middle East Technical University Inonu Bulvari, 06800, Ankara, Turkey

Abstract

Particle Swarm Optimization (*PSO*) is a member of the nature inspired algorithms. Its ability to solve many complex search problems efficiently and accurately has made it an interesting research area. In this study, we model Distributed Database Query Optimization problem as a *Bare Bones PSO* and develop a set of canonical and hybrid *PSO* algorithms. To the best of our knowledge, this is the first time that *Bare Bones PSO* is being used for solving this problem. We explore and evaluate the capabilities of *PSO* against *Iterative Dynamic Programming*, and a *Genetic Algorithm*. We experimentally show that *PSO* algorithms are able to find near-optimal solutions efficiently.

Key words: Particle swarm intelligence, Distributed database, Query optimization, Bare Bones

1 Introduction

Particle Swarm Optimization (*PSO*) is a stochastic optimization technique developed by Kennedy and Eberhart in 1995 [1]. In *PSO*, each member of the swarm is moved through the problem space by two forces. One force attracts it with random magnitude towards the best location so far encountered by the particle, and the other attracts it with random magnitude towards the best location encountered by any member of the swarm. In this paper, we study the NP-Hard *Distributed Database (DDB)* query optimization problem with *PSO* algorithms. Much research has been done on *DDB Management Systems* [2], mostly focusing on the optimization of single queries. Query optimizers are compared according to the amount of main memory consumed while searching for a good query plan and the limited optimization time are the major factors of a query optimizer. Since this is an NP-hard problem [3], query optimizers make tradeoffs between the optimization time and the execution plan quality. In this study, we present and evaluate a set of *Standard* and *Hybrid PSO* algorithms. Developed *Hybrid PSO* algorithms borrow the idea of mutation from the evolutionary strategies [4]. *PSO* algorithms efficiently work for the optimization of *DDB* queries and produce query execution plans that are comparable to those produced by the genetic

algorithms. We performed the experiments with several synthetic databases derived from the TPC-H benchmark (<http://www.tpc.org/tpch/>). SQL statements include multi-way chain-join queries up to 12 ways. In Section 2, an overview of the related work is presented. Section 3 describes *DDB* query optimization problem. Section 4 examines fundamental algorithms that inspired us to use *PSO* for solving this problem and gives our model. Section 5 presents the results of our experiments to measure the performance of the developed *PSO* algorithms and compares them with *Iterative Dynamic Programming* and *Genetic Algorithms*. Section 6 gives our concluding remarks.

2 Previous Works

Many algorithms have been proposed for the query optimization in *DDBs*. *Distributed INGRES*, *R**, and *SDD-1* are the early examples of these algorithms [5]. Query optimization algorithms can be classified as exhaustive, approximation, and randomized algorithms. Dynamic Programming (*DP*) and A* are exhaustive algorithms that are widely used in centralized and distributed *DBMSs*. It is possible to use approximation algorithms that have polynomial time and space complexities, however, optimal solutions are not guaranteed by such algorithms. Another alternative is to use randomized algorithms that search the solution space by randomly generating solutions. Nana proposed a Swarm Intelligence constructive algorithm called "Ant Colony Optimization Algorithm

Email addresses: tansel@ceng.metu.edu.tr (Tansel Dokeroglu), tosun@ceng.metu.edu.tr (Umut Tosun), cosar@ceng.metu.edu.tr (Ahmet Cosar).

for Multi-Join Query Optimization” [6], where a set of artificial ants build feasible query execution plans. Ant colony optimization metaheuristics [7] and genetic algorithms are the other algorithms that have been successfully applied for *DDB* query optimization [18,19]. *PSO* has been widely used to solve many different optimization problems since it was proposed by Kennedy [8]. Angeline developed one of the first hybridized *PSO* algorithms with evolutionary paradigm. In hybrid *PSO*, a selection is applied to the population of particles, new particles are produced with crossover and mutation, and bad particles are swapped with them. Our test results show that hybrid *PSO* benefits from these ideas.

3 Querying distributed databases

A *DDB* is a collection of data which resides on more than one database server with data processing capability connected via a reliable communication network. *DDBMS* is a software system that permits the management of a *DDB* and makes the distribution transparent to users. In a *DDBMS*, data is stored at a number of sites in a computer network. Each site is assumed to logically consist of a single processor, with all resources included in a single computer system. The computers are interconnected with a computer network rather than by a multiprocessor switch because computer networks are proven to be scalable and cost efficient. Sites have their own operating systems and operate independently so that they are loosely interconnected. *DDBMSs* support data fragmentation, where a relation in a database is divided into pieces called fragments for physical storage. In horizontal fragmentation, splitting creates subsets of tuples. In vertical splitting, a projection of relation is performed using subsets of attributes on the complete relation. A database is called "partitioned" if these horizontal and/or vertical fragments are placed at different sites. The problem of determining good evaluation strategies for join expressions has first been addressed by System R [3]. The input of the optimization problem is given as a query graph. A query graph consists of all the relations that are to be joined as its leaf nodes. Non-leaf nodes indicate the joins and can be labelled with the join selectivity which denotes the ratio of the included tuples to the total tuples. The search space or solution space is the set of all possible evaluation plans that generates the desired result. A solution is described by the corresponding processing tree for evaluating a join expression. The processing tree, which we will use in the next section to show the join orders of the execution plans is a binary tree. Its leaves correspond to base relations and inner nodes correspond to join operations. Edges indicate the flow of partial results from the leaves to the root of the tree. Each evaluation plan has a cost (CPU + disk I/O time). The aim of the optimization is to find the evaluation plan with the lowest possible cost [14]. Generally, the solution space is defined as the set of all processing trees that compute the result of the join

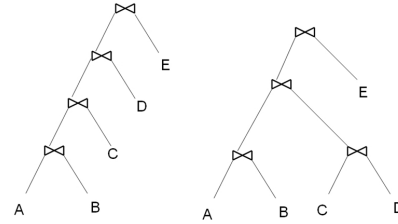


Fig. 1. Left-deep and bushy trees.

expression and that contain each base relation exactly once. While leaves of the processing trees consist of the base relations, inner nodes match joined results of their child nodes. If the inner relation (right child) of each join is a base relation, this type of processing tree is called left-deep. There are $n!$ ways to allocate n base relations to the leaves of the tree for processing. If there is no restriction on the shape of the processing tree, it is called bushy [12] (Figure 1).

4 *PSO* model for the *DDB* query optimization

4.1 Overview of the examined Particle Swarm Optimization Algorithms

The discipline dealing with artificial and natural systems of individuals that coordinate using decentralized control and self-organization is called *Swarm Intelligence (SI)*. It focuses on the collective behaviors resulting from the local interactions of the individuals with each other and their environment. Ant colonies, termites, flocks of birds, and herds of land animals are examples of systems studied by *SI*. *PSO* is inspired by the social behavior of birds flocking or fish schooling [8]. At first, *PSO* was a simulation but later developed to be a powerful optimization method. Potential solutions (particles) move through the problem space by following the paths of its successful neighbors. Each individual keeps track of its best fitness value and that of its neighbors' and uses this information to determine its future direction and velocity. As a result, the swarm, like a flock of birds foraging for food, moves towards an optimum value of the function. Each particle has a position and a velocity. The position of the particle is shown with a vector and this vector represents a possible solution for the problem instance. At each iteration, the new position of a particle is calculated depending on its position and a velocity. If this new particle's solution is better than the previous ones, it is kept as the local best particle. If its fitness value is better than global best fitness value, it is also kept as the global best particle. Equations 1 and 2 show the velocity and the new position calculation of the particles respectively. The canonical *PSO* is shown in Algorithm 1.

$$V_{id(t+1)} = \omega V_{id(t)} + C_1 \phi_1 (p_{d(t)} - x_{id(t)}) + C_2 \phi_2 (g_{d(t)} - x_{id(t)}) \quad (1)$$

Algorithm 1 Canonical *PSO* algorithm.

Input: Position and velocity of the particles
Output: Position of the approximate global optima
Begin
while terminating condition is not reached **do do**
 for $i = 1$ to number of particles **do do**
 Begin
 Evaluate the fitness;
 Update $p(t)$ and $g(t)$;
 Update velocity of the particle;
 Update the position of the particle;
 end for
end while

$$x_{id(t+1)} = x_{id(t)} + V_{id(t+1)} \quad (2)$$

The canonical *PSO* has a number of parameters that need to be determined (Table 1). ϕ_1 and ϕ_2 are acceleration coefficients that determine the random forces in the direction of the local best (p_d) and and global best (g_d). The stability of *PSO* can be adjusted by changing the acceleration coefficients. The speed must be set carefully otherwise, an uncontrolled parameter may cause undesirable movements in the search space of the problem. To prevent this side effect, a limiting the velocity can be used $[-V_{max}, +V_{max}]$.

Table 1

Parameters of *PSO*

Symbol	Parameter
χ_{id}	Position of particle i at dimension d
V_{id}	Velocity of particle i at dimension d
ω	Inertia weight
C_1	Importance of local best
C_2	Importance of neighborhood best
ϕ_1	Acceleration coefficient of local best
ϕ_2	Acceleration coefficient of global best
p_d	Local best value of the particle at dimension d
g_d	Global best value of particles at dimension d

Another parameter is inertia weight ω that is designed to control the scope of the search space and reduce the importance of V_{max} value. Higher values of ω represents a low viscosity medium and performs a more extensive search. It is a frequently used approach to start with a higher value of ω and decrease it gradually during the optimization process. When the *PSO* algorithm is run without any constraints in the velocity values, in a few iterations there happen unacceptable increases in discovered solution values [13]. Clerc and Kennedy have proposed a new strategy for the placement of constriction coefficients [9]. Many ways are mentioned to implement constriction coefficient parameters. The simplest one of is given in Equations 3 and 4. With this model, there is no need to use V_{max} and constricted particles can converge without it. An even better approach is setting a limit to V_{max} as χ_{max} . This rule of thumb pro-

vides problem-independent parameters. This method is called canonical *PSO*. *PSO* with constriction is equivalent to *PSO* with inertia. Equations 1 and 3 can be converted to each other by mapping χ to ω and ϕ_i to $\chi\phi_i$. In canonical *PSO*, particles are influenced by their own best values and the global best value. A particle is not affected by other particles. Kennedy and Mendes have improved the way particles react with each other [10]. Instead of particles interacting with only its previous best or the best of its neighbors, particles can influence each other. This model is called Fully Informed Particle Swarm (*FIPS*) [1]. The formula of *FIPS* is given in Equation 5. K_i is the number of neighbors for particle i , and nbr_n is i 's, n th neighbor. When $K_i=2$, *FIPS* acts like the standard *PSO*. *FIPS* is proved to perform better than the standard *PSO*. Its success depends on the population topology that Kennedy proposed as a new method instead of velocity updating to decide the new locations of the particles [11]. The velocity update rule is decided with Gaussian distribution of mean as given in Equation 6 and standard deviation as given in Equation 7. *Bare Bones* is another *PSO* algorithm inspired by a single particle moving in one dimension under the influence of fixed \vec{p}_i and \vec{p}_g . This distribution resembles a Gaussian function centered at $(\vec{p}_i + \vec{p}_g)/2$ (Equation 8).

$$V_{id(t+1)} = \chi(V_{id(t)} + C_1\phi_1(p_{d(t)} - x_{id(t)}) + C_2\phi_2(g_{d(t)} - x_{id(t)})) \quad (3)$$

$$\chi = \frac{2}{\phi - 2 + \sqrt{\phi^2 - 4\phi}} \quad (4)$$

$$V_{id}(t+1) = \chi(V_{id}(t) + \frac{1}{K_i} \sum_{n=i}^{K_i} \vec{U}(0, \phi_1) \times (\vec{p}_{nbr_n} - \vec{x}_i)) \quad (5)$$

$$\mu_{ij} = \frac{(p_i + g_i)}{2} \quad (6)$$

$$\sigma_{ij} = |p_i + g_i| \quad (7)$$

$$x_{ij} = N(\mu_{ij} + \sigma_{ij}) \quad (8)$$

TRIBES is a parameter-free variant *PSO* model. Its topology and the size of the population evolve over time. The population is divided into subpopulations and these

subpopulations may eliminate the worst performing particles or add new particles [14]. In Clerc’s experiments, population is evaluated at every (number of links / 2) iteration. Bird-flocking simulations were the first developed population topologies [13]. Trajectory of each bird is decided by the nearby location of the others. Earlier *PSO* topologies used this proximity. This approach has inefficient convergence results. Global best (*Gbest*) is a static topology where the best particle affects the location of the current particle. In static topologies neighbors and neighborhoods do not change. In static Local best (*Lbest*) topology, each individual is connected to adjacent members in the population in a ring lattice. This population structure allows parallel search [10]. Mendes proposed a stochastic average of all neighbors’ previous best positions with *FIPS* [15]. Although static topologies are used, dynamic topologies can be efficient and many researchers have reported results supporting this idea. Clerc designed a dynamic population parameter-free *PSO*, *TRIBES*. The size of the population evolves over time in parallel with the performance. Each subpopulation maintains its order and structure. Subpopulations may change their weakest particles and create random members. This is a hybrid model employing the genetic algorithms. This property improves the possibility of producing better solutions.

4.2 Solution Model for the DDB query Optimization

At first, we have modeled the *DDB* query optimization problem with the velocity based *PSO* algorithms. Most of these *PSO* models are based on real-numbers. Our particles needed to move on a space of integers. Setting the right parameters for these algorithms was not easy and rounding the real numbers usually caused search space limits to be exceeded. Our velocity tests with canonical *PSO* and *FIPS* did not result in efficient ways to move particles. Most of the time, they exceeded the search space limits and reached the boundaries of the particles, prematurely converging on a global best position without finding an optimal solution. *TRIBES* is the other *PSO* algorithm we have not gained much performance. The most efficient algorithms we have developed can be mainly classified as *standard* and *hybrid Bare Bones PSO* algorithms, which are implemented by using *LBest*&*GBest* neighborhood communication structures. *LBest* is the best fitness value permutation in the history of that particle and *GBest* is the best particle in the population. Mutation is the only operator of evolutionary paradigm we have applied to the selected particles. The algorithm computes the average of the neighboring particles and uses it as an attractor. In the *LBest*&*GBest* communication model, the best past value of the particle and the global best value of the population are used to decide a new location for the particle. Each particle is represented with a sequence of integers. For example (2,4,6,3,1) represents a 6-way chain join. This sequence is a 5 dimensional space, where particles move. Each di-

mension is limited between $[1, M]$ (M is the total number of the sites where the joins can be performed). It is possible to check all the evaluation alternatives of the join operations with this model. The first join of a left-deep tree will be done at site 2, and the results will be input to the second join which will be executed at site 4, and so on. The multi-dimensional space of the particles and how the order of the relations are applied to this space is explained in Section 5.2. The structure of the population topologies we have applied are static and there are limits for the minimum and maximum positions of the particle. For example, we use values between $[1,6]$ for a 6-way join where 6 is the total number of the sites that the join can be done. For particle (1,5,1,6,4), if we select the dimension 3, (particle position at this dimension is: 1) and if the local best and global best values are 2 and 4 respectively, then the new value of the particle at this dimension is evaluated to be 3. We have moved the particle in the selected dimension at each iteration. In Figure 2, the placement mechanism of a particle can be seen in a two-dimensional space. Using this mechanism, a velocity-independent model has been used to change the positions of particles. In hybrid algorithms, we have applied mutation to the selected particle of the population to replace the particle having the worst response-time in the population. The hybrid *PSO* algorithm is listed in Algorithm 2.

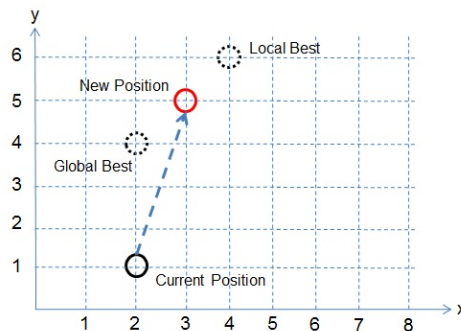


Fig. 2. Calculating the new position of a particle using "local best" and "global best" values.

5 Experimental setup and results

5.1 Experimental environment

We have evaluated our algorithms through a number of experiments using a 2.21 GHz AMD Athlon (TM) 64x2 dual processor with 2 GB RAM and MS Windows 7 (TM) operating system. Each relation is assigned to a single site in the distributed database environment. The underlying network is assumed to have a fully connected graph topology, thus no multi-hop transmissions and store-and-forward delays are needed. There is no fragmentation or replication of relations at any site. Different databases are generated with the TPC-H benchmark (<http://www.tpc.org/tpch/>). The database schemas allow up to 12-way chain queries. The relations are as-

sumed to have different cardinalities and the referential integrity is guaranteed to be satisfied by all relations in the schema. Each processing node has 102 buffers (each buffer is one page), and page size is 10,240 bytes, disk I/O time is 10ms (per page), available memory is sufficient to perform all join operations in main memory and each table is loaded into memory only once. The implementation language is C++. We limited our SQL statements to multi-way chain equijoin queries with simple selection predicates. Cross Product operations are not enumerated by the algorithms. An example 4-way chain-join can be given as:

```
SELECT K.NAME FROM K, L, M, N
WHERE(K⋈ L ⋈ M ⋈ N )
AND N.RETAILPRICE < 1,000
```

Algorithm 2 Hybrid *PSO* algorithm.

```
Input: Position and velocity of the particles
Output: Position of the approximate global optima
Iteration number:=0;
while termination condition is not reached do
  if iteration number is even then
    for for i = 1 to number of particles do
      Apply mutation operator to the particle(i)
      Evaluate the fitness of the particle(i);
      Update p(t) and g(t);
    end for
  end if
  if iteration number is odd then
    for for i = 1 to number of particles do
      Find the new place of the particle (i)
      Evaluate the fitness of the particle(i);
      Update p(t) and g(t);
    end for
    Iteration number++;
  end if
end while
```

5.2 Comparing the results of the algorithms

We have evaluated the performance of our algorithms by comparing the results of *Iterative Dynamic Programming (IDP)* and a Genetic Algorithm (*SGA₂*). *IDP* is a constructive approximation algorithm [20]. *SGA₂* is an improved variant of the *SGA* algorithm given in [8]. Detailed parameters of the *SGA₂* can be found in Table 2. Population size in *SGA₂* is 100 and the members of the initial population are non-redundant. At each iteration of the algorithm, we implement half-population size crossovers. Mutations are randomly generated and crossovers are one-point. Survival mechanism eliminates the worst plans from the population. The termination condition is decided as having less than a threshold value improvement in the average quality of the population in a new generation. For up to 6-way joins, a population size of 500 and having at least 5% improvement in each iteration were sufficient to find an optimal solution. To

measure the statistical confidence of *SGA₂*, we repeated each *SGA₂* execution 20 times and calculated standard deviations of discovered solution values and validated that all results are within 2% of the averages.



Fig. 3. Chromosome structure for a distributed 4-way chain-join using three sites to complete the SQL statement.

Each chromosome in *SGA₂* has genes that represent sites of the individual join operations. The chromosome structure of the SQL given in section 5.1 can be seen in Figure 3. The execution order of the joins (we consider only left-deep trees) show that relation K and L are joined at site 2 to obtain $(K \bowtie L)$, then, this result and relation M are both shipped to site 1 to produce $((K \bowtie L) \bowtie M)$. Finally, relation N is shipped to site 1 for $(K \bowtie L \bowtie M \bowtie N)$. With n sites, there are n^n different ways to organize the genes of a chromosome. Possible different evaluation orders of 4-way join of $(K \bowtie L \bowtie M \bowtie N)$ are shown in Table 3. The number of the relation orders reduces by the commutativity property $(K \bowtie L = L \bowtie K)$ of the join operation. These orders of the join do not have any Cross Products. The number of the possible orders increases as the number of sites gets larger. In the experiments, each chromosome is evaluated with its all possible left-deep processing tree orders. In this part of the study, we have evaluated the algorithms of *TRIBES-PSO* with a static population. Our aim was to evaluate whether *TRIBES-PSO* algorithms are better than standard and hybrid *PSO* algorithms and can be used for comparison or not. In *TRIBES-PSO* algorithms, members of each subpopulation communicate within their communities. They don't have any information about what the other subpopulations are doing. This property prevents all the subpopulations from moving towards a local optimum found by one of the others. We have concluded from the results of our experiments that when subpopulations are used in *PSO* algorithms, both the running times of the algorithm and the quality of the execution plans decrease. The comparison of *TRIBES* algorithms with the standard *PSO* algorithm can be seen in Figure 4.

Table 2
Parameters of *SGA₂* algorithm.

Parameter	Model
Initialization:	Random
Representation:	Each gene gives join site number
Population Size:	100
Mutation:	1 %
Crossover:	One-point
Survival selection:	Tournament
Termination Condition:	Less than 5% Improvement

Table 3

All possible left-deep tree orders of (K \bowtie L \bowtie M \bowtie N) 4-way join. (with $4! = 24$ different plans) search space that is reduced to 4 by the commutativity property of the non-cross product join operation)

1. K \bowtie L \bowtie M \bowtie N
2. L \bowtie M \bowtie K \bowtie N
3. L \bowtie M \bowtie N \bowtie K
4. M \bowtie N \bowtie L \bowtie K

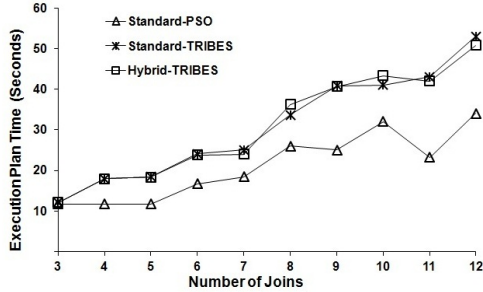


Fig. 4. Execution plan quality of TRIBES-PSO algorithms.

5.3 Optimization times and execution plan quality of PSO algorithms

We have performed our experiments using our new set of PSO algorithms, IDP, and SGA₂ algorithms. We have extended our experiments up to 12-way chain-join queries. IDP was able to find optimal 12-way chain-join execution plans in reasonable running times. We compare the query execution plans of IDP with other algorithms. Figures 5 and 6 show the optimization times and execution plan qualities of standard and hybrid PSO algorithms respectively. Running times of both algorithms are very close to each other and both can produce execution plans with the same quality. As the number of joins increases, the optimization time also increases, but still it is possible to produce good quality execution plans in reasonable times.

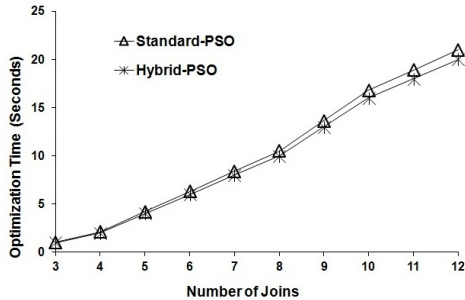


Fig. 5. Optimization times of PSO algorithms.

Figures 7 and 8 summarize the optimization times and the query execution plan qualities of the evaluated algorithms. Standard and hybrid PSO, SGA₂, and IDP algorithms have almost the same optimization times. Population sizes and the required improvement percentage in

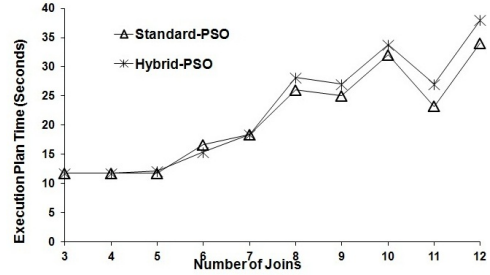


Fig. 6. Execution plan quality of PSO algorithms.

the average response time of the particles in population at each iteration are the main factors of the optimization time of the algorithms. We use a 5% improvement requirement at each iteration for all the algorithms. It is observed that standard and hybrid PSO algorithms can explore the search space as effectively as SGA₂ algorithm with the same parameter settings. Since the quality of execution plans found by TRIBES-PSO algorithms are much worse than the other algorithms, they are not included in the final diagrams.

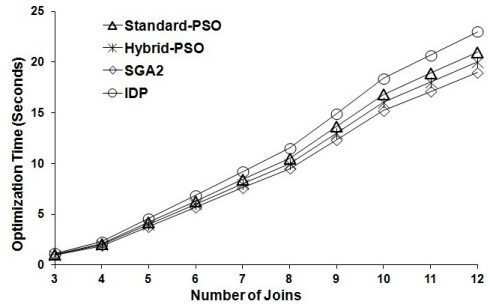


Fig. 7. Optimization times of all algorithms.

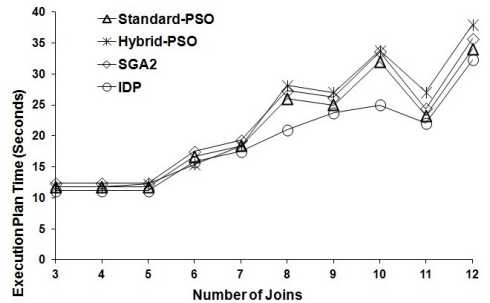


Fig. 8. Execution plan qualities of all algorithms.

The quality of query execution plans of the algorithms are evaluated not only from the perspective of discovered plan quality but also from the perspective of optimization times. Best execution plans are produced by IDP algorithm, but its optimization time is not as efficient as other algorithms. The optimization times of the algorithms can be seen in Figure 7 as the number of joins increases. In most of the problem instances, SGA₂ algorithm has the smallest optimization time. Standard and hybrid PSO algorithms can find the same quality execution plans like those of SGA₂'s (Figure 8).

6 Conclusions and future work

In this study, we present a new set of *Particle Swarm Intelligence (PSO)* algorithms for the optimization of distributed database queries. *PSO* algorithms can provide high quality solutions like those of genetic algorithms. In the algorithms, particles representing the solutions are moved according to a probability distribution rather than by manipulating the velocity parameter. This mechanism ensures a simple and effective way to explore the search space. As new and better global and local solutions are discovered, *PSO* algorithms keep searching and can produce near-optimal quality execution plans. The execution times and the quality of produced execution plans can be adjusted by varying the termination conditions and the parameter settings of the *PSO* algorithms. As future work, we are planning to apply *PSO* algorithms to multiple query optimization and the design of distributed databases.

References

- [1] Poli R, Kennedy J, Blackwell T. Particle Swarm Optimization, *Swarm Intelligence*, 2007; Vol.1, pp. 33-57.
- [2] Kossmann D. The State of the Art in Distributed Query Optimization, *ACM Computing Surveys*. 2000; vol.32, pp. 422-469.
- [3] Ibaraki T, Kameda T. On the optimal nesting order for computing N-relational joins. *ACM Trans. Database Syst.* 1984; 9, 3, pp. 482-502.
- [4] Miranda V, Fonseca N. New evolutionary particle swarm algorithm (EPSO) applied to voltage/VAR control. In *Proceedings of the 14th power systems computation conference*. 2002; Session 21, Paper 5, pp. 1-6, Seville, Spain.
- [5] Ozsu MT, Valduriez P. *Principles of Distributed Database Systems*, 3rd Edition, 2011; pp. 245-293.
- [6] Nana L, Yujuan L, Yongfeng D, Junhua G. Application of Ant Colony Optimization Algorithm to Multi-join Query Optimization, *LNCS*, 2008; vol.5370,pp.189-197.
- [7] Dokeroglu T, Cosar A. Dynamic Programming with Ant Colony Optimization Metaheuristic for The Optimization of Distributed Database Queries, in *Proc. of the 26th Int. Sym. On Computer and Information Sciences*, 2011; London, UK.
- [8] Kennedy J, Eberhart R C. Particle swarm optimization. In *Proceedings of the IEEE international conference on neural networks IV*, 1995; pp. 1942-1948.
- [9] Clerc M, Kennedy J. The particle swarm explosion, stability, and convergence in a multidimensional complex space. *IEEE Transaction on Evolutionary Computation*, 2002; 6 (1), pp. 58-73.
- [10] Kennedy J, Mendes R. Population structure and particle swarm performance. In *Proceedings of the IEEE congress on evolutionary computation*. 2002; pp. 1671-1676, Honolulu, HI. Piscataway: IEEE.
- [11] Steinbrunn M, Guido M, Kemper A. Optimizing Join Orders, Technical Report MIP9307, 1993; Faculty of Mathematic, University of Passau, Germany.
- [12] Kennedy J, Eberhart R, Shi Y. *Swarm Intelligence*, 2001; Morgan Kaufmann Academic Press.
- [13] Kennedy J. The behavior of particles. In V.W. Porto, N. Saravanan, *LNCS Evolutionary Programming VII: 1998*; pp. 581-589. San Diego, CA. Berlin: Springer.
- [14] Clerc M. *Particle swarm optimization*. London: ISTE 2006.
- [15] Mendes R. Population topologies and their influence in particle swarm performance. Ph.D. thesis, 2004; Departamento de Informatica, Escola de Engenharia, Universidade do Minho.
- [16] Toroslu IH, Cosar A. Dynamic programming solution for multiple query optimization problem. *Information Processing Letters*. 2004; Vol 92, pp. 149-155.
- [17] Bayir MA, Toroslu IH, Cosar A. Genetic Algorithm for the Multiple-Query Optimization Problem. *IEEE Transactions on Systems, Man, and Cybernetics, Part C*, 2007; vol.37 pp. 147-153.
- [18] Sevinc E, Cosar A. An Evolutionary Genetic Algorithm for Optimization of Distributed Database Queries, *The Computer Journal*, 2011; vol.54, Issue 5, pp. 717-725.
- [19] Dokeroglu T, Tosun U, Cosar A. Parallel Mutation Operator for the Quadratic Assignment Problem, *Proceedings of WIVACE*, 2012; Italian Workshop on Artificial Life and Evolutionary Computation.
- [20] Kossmann D, and Stocker K. Iterative Dynamic Programming: a New Class of Query Optimization Algorithms. *ACM Transactions on Database Systems*, vol.25, issue 1, 2000, 43-82.

Realization of Direct Series of Discussions in the Face of Conflicting Set of Production Rules

Roman SAMKHARADZE

Department of Computer
Engineering
Georgian Technical University
Tbilisi, Georgia.
samkharadze.roman@gmail.com

Davit CHIKOVANI

Faculty of Computer
Technologies and Engineering
International Black Sea
University
Tbilisi, Georgia.
ibsu.geo@gmail.com

Lia GACHECHILADZE

Department of Computer
Engineering
Georgian Technical University
Tbilisi, Georgia.
gachechiladze.lia@gmail.com

Abstract—The article considers the algorithm of operation of interpreter of expert system in the face of conflicting set of production rules. Algorithms of solution of conflicts among the production rules are developed in the process of making logical conclusions by the example of management of normal daily modes of power system.

Keywords: *expert system, power system, technologist, production rules*

I. INTRODUCTION

Nowadays, solution of conflicts among production rules is the important problem. On the one hand, this problem complicates the effective management of complex objects and processes, on the other hand, making of correct decisions is harder. The article presents one of the approaches to solve this problem.

The main content

The expert system [1-3] of the management of daily modes of Georgian power system has the block of logical conclusions. The main part of this block is the interpreter realizing the direct series of discussions. Interpreter also analyzes the **Condition** part of a rule. If the condition is met,

then interpreter acts according to the **Conclusion** part of the rule. Algorithm of operation of interpreter can be presented as following:

1. Initial condition is defined. Magnitudes are given to the operating variables;
2. Variables and conditions are carried in the series of logical output's variables;
3. Magnitudes of the variables of the condition are carried in the list of variables;
4. The variable is searched in the list of variables, name of which is in the beginning of series of logical output's variables;
5. If such variable is found, then in the reference of condition variables, number of rule and number 1 are inserted;
6. If variable isn't found, then we transfer to the 10th difference.
7. The corresponding magnitudes are given to the non-initialized variables of the **Condition** part of the discovered rule. The names of variables are in the list of condition variables;

8. Every condition of the rule is checked and if these conditions are true, then **Then** part of the rule is processed;
9. The corresponding magnitudes are given to the variables of **Then** part of the rule and they are placed at the end of logical output's series;
10. If the variable is at the beginning of the list of logical conclusion's variables and isn't at the **Condition** part of the rule, then it is derived from this series;
11. Process of logical output is ended, when series of variables of logical output is empty;
12. If series of variables of logical output isn't empty, then we transfer to the 4th difference.

Let's consider the example of operation of interpreter. Considered section of knowledge base has the following form:

1. If at hour t $\psi(t) > 0$ & $N_{TPS_j}(t) > N_{TPS_j}^{\min}$, then $N_{HPS_i}(t) = N_{HPS_i}^{\min}$

2. If at hour t $\psi(t) > 0$ & $N_{TPS_j}(t) > N_{TPS_j}^{\min}$, then $N_{HPS_i}(t) = 0$

3. If at hour t $\psi(t) > 0$ & $N_{TPS_j}(t) > N_{TPS_j}^{\min}$, then $N_{TPS_j}(t) = N_{TPS_j}^{\min}$

4. If at hour t $\psi(t) > 0$ & $N_{TPS_j}(t) = N_{TPS_j}^{\min}$, then dispatcher of power system begins to manage.

For the clarity, scheme of operation of interpreter is presented in the form of table. Names of variables are listed in the first table.

The first rule is priority. Therefore, it's processed firstly. Then the second rule is considered and at last we have the third rule. It means that if hydroelectric power station has the surplus of water, then at hour t it must generate the

minimal power $N_{HPS_i}^{\min}$. If at this hour generation of hydroelectric power station is $N_{HPS_i}^{\min}$, then it should be turned off. If such activities didn't give the result, then the power station must generate the minimal power at the given hour.

Knowledge base includes list of condition variables, i.e. list of variables included in the **Condition** part of each rule (Table 2). Let's assume, at the fourth daily hour we have the positive imbalance with magnitude 90 megawatt and power of i th hydroelectric power station is 600 megawatt. Variables $N_{HPS_i}^{\min}$, $\psi(t)$ and t are in the **Condition** part of the rule. Therefore, they are carried in the list of logical output's variables (Table 3)

Table 1. Names of Variables

Name of variable	Meaning
t	daily hour
$\psi(t)$	imbalance of HPS at hour t
$N_{HPS_i}(t)$	power of i th HPS at hour t
$N_{TPS_j}(t)$	power of j th TPS at hour t
$N_{HPS_i}^{\min}$	minimal power of i th HPS
$N_{TPS_j}^{\min}$	minimal power of j th TPS

Table 2. Condition Variables

List of Condition Variables	Sign of Initialization	Meaning
t	non-initialized	
$\psi(t)$	non-initialized	
$N_{HPS_i}(t)$	non-initialized	
$N_{TPS_j}(t)$	non-initialized	
$N_{HPS_i}^{\min}$	non-initialized	
$N_{TPS_j}^{\min}$	non-initialized	

Table 3. Variables of Logical Output

$\psi(t)$
t
N_{HPSi}^{mn}

Table 4. List of Variables of Logical Output

$\psi(t)$	← beginning
t	
N_{HPSi}^{mn}	
	← end

Table 5. Variables of Logical Output

$N_{HPSi}(t)$

Table 6. Condition Variables

List of Condition Variables	Sign of Initialization	Magnitude
t	initialized	4
$\psi(t)$	initialized	90
$N_{HPSi}(t)$	initialized	60
$N_{HPSi}(t)$	initialized	750
N_{HPSi}^{mn}	initialized	70
N_{TPSj}^{mn}	initialized	700

As the above-mentioned suggests, the **Condition** part of the first rule is fulfilled. Therefore, procession of the second sub-conclusion of the **Conclusion** part begins. Current variable is $N_{HPSi}(t)$, which has zero magnitude. This variable is carried in the series of variables of logical output (Table 4). After every rule including variables t, $\psi(t)$ and $N_{HPSi}(t)$ is

checked, these rules are carried out from the series of logical output's variables (Table 5).

Procession of variable $N_{HPSi}(t)$ is similar to variable $N_{TPSj}(t)$. As soon as series of logical output's variables is empty, direct series of discussions ends and the problem is solved.

List of variables includes magnitudes and signs of initialization of variables. At the beginning of the work the sign of initialization is "non-initialized" and the variables have empty magnitudes (Table 2). In the process of dialogue variables have magnitudes and the sign of initialization is „initialized“ (Table 6).

Variable referent includes information about the rules used by the expert system at the present time. Variable referent consists of number of rule and number of condition in the rule, because the **Condition** part of the rule may include several conditions. Referent is used to monitor the current situation in the process of discussion. The first rule includes the searchable variable. Therefore, referent references to the first rule (Fig. 1).

For the second case the same process repeats, but for the first sub-conclusion of the **Conclusion** part of the first rule. Results are stored to avoid the repeated calculations, if the initial data is the same.

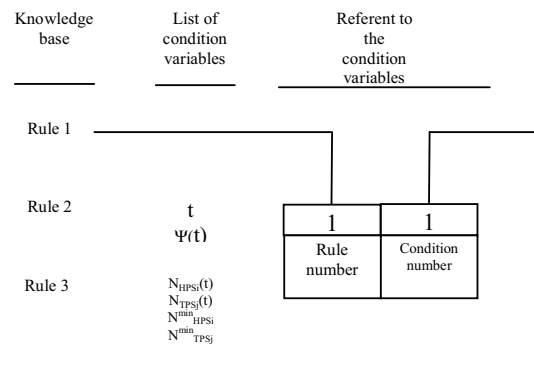


Fig. 1.

Expert system stores the results. If the initial data is the same, it makes the corresponding decisions without repeated calculations. In the face of conflicting set of production rules, expert system utilizes heuristic knowledge. According to one of the heuristics, load of hydroelectric and thermal power stations should begin from the peak hour and then decrease. Therefore, hours should be categorized according to the reduced consumption. According to one more heuristics, hydroelectric power stations should be categorized on the basis of ability of regulation and these stations should be loaded correspondingly. It means that first of all hydroelectric power stations with good ability of regulation should be loaded. This approach and utilization of additional information frequently solves conflicts among the production rules efficiently.

Conclusion

In the article algorithm of operation of interpreter of expert system is developed in the face of conflicting set of production rules. Algorithms of logical conclusions are developed and the corresponding example is given.

Reference

1. R. Samkharadze, D. Chikovani. Expert System in administering the Georgian power system. Georgian Technical University. The International Scientific Conference Devoted to the 80th Anniversary of Academician I.V. Prangishvili «Information and computer Technologies, Modelling, Control» Tbilisi, Georgia, 2010. ISBN: 978-9941-14-855-2, pages: 162-163
2. R. Samkharadze, D. Chikovani. Ways and means of a resolution of conflicts between production rules. Transactions Automated Control Systems, Journal #2(11). Tbilisi, 2011. ISSN: 1512 – 3979, pages: 49-56
3. R. Samkharadze, D. Chikovani. Development of Structure of Expert System on the Basis of Solution of Conflicts among Production Rules. Scientific Journal of International Black Sea University. Volume 4, Issue 2, Tbilisi, 2010. ISSN 1512-3731. p. 157-167.

Summary

In the article algorithm of solution of conflicts among production rules is developed in the process of making logical conclusions. Efficiency of algorithm is shown on the example of management of daily modes of power system. This approach hastens the process of making correct and effective decisions.

Towards Remote Security Monitoring in Cloud Services Utilizing Security Metrics

Reijo M. Savola

VTT Technical Research Centre of Finland

Oulu, Finland

address: VTT, Kaitoväylä 1, FIN-90570 Oulu, Finland

phone: +358 40 569 6380

fax: +358 20 722 2320

email: reijo.savola@vtt.fi

Jukka Ahola

VTT Technical Research Centre of Finland

Oulu, Finland

address: VTT, Kaitoväylä 1, FIN-90570 Oulu, Finland

phone: +358 40 076 4010

fax: +358 20 722 2320

email: jukka.ahola@vtt.fi

Abstract— Large amounts of business-critical data are transferred, processed and stored in cloud services, raising concerns about their security level. Adequate security management of cloud services is vital to their success. Systematically developed and maintained security metrics can be used to offer evidence of the security effectiveness of cloud services. We propose a metrics based approach for remote security correctness monitoring in the Cloud. The approach was investigated by building a monitoring system within an experimental cloud system set-up. Moreover, we discuss how risk-driven security metrics modeling based on the decomposition of security objectives is used to manage monitoring activities.

This paper has neither been published before nor currently is being submitted elsewhere.

Reijo M. Savola and Jukka Ahola

Towards Remote Security Monitoring in Cloud Services Utilizing Security Metrics

Reijo M. Savola

VTT Technical Research Centre of Finland
Oulu, Finland

address: VTT, Kaitoväylä 1, FIN-90570 Oulu, Finland
phone: +358 40 569 6380
fax: +358 20 722 2320
email: reijo.savola@vtt.fi

Jukka Ahola

VTT Technical Research Centre of Finland
Oulu, Finland

address: VTT, Kaitoväylä 1, FIN-90570 Oulu, Finland
phone: +358 40 076 4010
fax: +358 20 722 2320
email: jukka.ahola@vtt.fi

Abstract— Large amounts of business-critical data are transferred, processed and stored in cloud services, raising concerns about their security level. Adequate security management of cloud services is vital to their success. Systematically developed and maintained security metrics can be used to offer evidence of the security effectiveness of cloud services. We propose a metrics based approach for remote security correctness monitoring in the Cloud. The approach was investigated by building a monitoring system within an experimental cloud system set-up. Moreover, we discuss how risk-driven security metrics modeling based on the decomposition of security objectives is used to manage monitoring activities.

Keywords—security; security metrics; cloud services; monitoring

I. INTRODUCTION

Cloud services allow rapid, flexible and economical outsourcing of computing tasks and have the potential to free organizations from the need to equip and manage substantial computing centers on their own. However, security concerns are emphasized in these services, affecting the role and extent of organizations' cloud service use especially in business-critical applications. Systematic, practical, sufficient and credible evidence of security effectiveness of cloud services are needed to enable trust and to have informed security management from the service user's perspective. Security transparency in the Cloud can be achieved by utilizing security metrics in SLAs (Service Level Agreements), security monitoring and security assurance activities.

High-quality RA (Risk Analysis) is crucial to effective security management. Based on prioritized RA results, it is possible to derive SOs (Security Objectives), and based on them, SCs (Security Controls). In systematic risk-driven security metrics development [1], SOs and SCs are used to set the reference level.

The main contribution of this study is an introduction of a security metrics based approach for remote security monitoring in the cloud. The approach is illustrated by a practical case example.

In this paper, we denote a customer or potential customer of a cloud service a *CSU (Cloud Service User)*, and the organization that offers the cloud service a *CSP (Cloud Service Provider)* [2].

II. BACKGROUND

In the following, we briefly discuss security risks of cloud services and development and maintenance of security metrics.

A. Security Threats in the Cloud

Cloud services are vulnerable to many generic security risks of software and telecommunication systems. Main SCs such as authentication, authorization, non-repudiation, availability, and integrity and confidentiality algorithms are needed in cloud services as in traditional ones. However, some risks (R) are more amplified compared to traditional service systems. These include [3]:

- (R1) Abuse and nefarious use of cloud computing. The relative anonymity of cloud services can be abused.
- (R2) Insecure interfaces. Several software interfaces can be utilized by the CSPs for CSU interaction of the service. There are concerns especially about APIs (Application Programming Interfaces).
- (R3) Malicious insiders. The threat of malicious insiders is amplified in the Cloud.
- (R4) Shared technology issues. Some underlying component parts of the cloud infrastructure not originally designed to that environment can potentially cause security problems.
- (R5) Data loss or leakage. The threat of data compromise increases in the Cloud due to the architectural and operational characteristics of the environment.
- (R6) Account or service hijacking. Service hijacking can have a large impact in the Cloud.
- (R7) Unknown risk profile. In the Cloud, from the perspective of CSU, tracking of security risks is lost without new approaches to it.

The above risks are obviously interdependent. For example, R6 can result from any other risks mentioned.

B. Security Metrics

The term (information) *security metrics* refers to security indicators of a System under Investigation (Sul) [4] – a technical system, product, service or organization. In addition to security monitoring, metrics can be used for to support risk

management, and security assurance, engineering and management.

In the high level, there are three main goals of security measurement: security effectiveness, security efficiency and security correctness. *Security effectiveness* goals aim at meeting the (SOs) Security Objectives in the SuI while the expectations for resilience in the use environment are satisfied [4, 5, 6]. In other words, it is the capability of the security controls in coping with the actual risks. *Security efficiency* goals aim at achieving sufficient security effectiveness from the perspective of practical resource, time, and cost constraints [4]. The focus of this paper is on security correctness. *Security correctness* can be defined as assurance that the security controls defined have been correctly implemented in the SuI, and the system, its components, the interfaces, and the processed data meet the security requirements [4, 5, 6]. Security correctness, which includes legal and standards compliance, is an important and concrete objective in practical security work, enabling security effectiveness goals.

III. SECURITY METRICS-DRIVEN MONITORING

A. Security Monitoring

Security monitoring activities most often concentrate on security correctness checking, such as:

- Correct configuration,
- Correct implementation and deployment of security controls,
- Regulation and legal compliance, and
- Standard compliance.

In many cases, security controls are defined correctly, and their implementation is adequate. However, the right kind of deployment tends to be a problem. In practice, the goal is to monitor the correctness by aiming at binary “yes” or “no” to be used as raw measurement outcomes utilized by metrics. In [7] and [8], we introduced reference architecture for building a general monitoring framework for security correctness monitoring. This approach consists of four layers: (i) at the bottom, the Base Measure Layer, (ii) next, above it, the Data Collection layer, (iii) the Measurement Control and Processing Layer, and, at the top, (iv) the Presentation, Evaluation, and Management Layer.

What makes cloud environments unique and challenging for security monitoring, is their dynamic nature of node server instances available at any time. From the security perspective, the potential security breach can occur in any of the available node instances in the cloud. Even one incident can be highly critical and it is easily hidden in the mists of the Cloud. Moreover, parts of the system or service that are unmanaged or managed by another party, are difficult to be monitored. Heterogeneity of components in cloud environments and the complexity arising from the wealth of runtime information are also remarkable challenges for monitoring [9].

In addition metrics-driven security monitoring, specific “security sensor” devices and applications, such as IDS/IPS

(Intrusion Detection and Prevention Systems), firewalls, and SIEM (Security Information and Event Management) systems can be used to offer evidence for monitoring and metrics management.

B. Metrics Model Development, Management and Visualization

In [1], we proposed a risk-driven security metrics development approach based on the decomposition of SOs.

The number of detailed-level metrics needed to offer sufficient security evidence can easily grow large. Especially in metrics hierarchies for security correctness, there are a wide variety of details to be addressed. An additional challenge is that the aggregation of measurement values has pitfalls: relying blindly on an aggregated value can result in loss of important information and can lead to a false sense of security. Visualization can be used to increase the manageability of SMMs. In [10], we introduced a modeling and visualization tool called the Metrics Visualization System, or MVS, for the management of hierarchical security metrics and measurements. In the MVS SMM, the basic building block is an SMN (Security Metrics Node). Hierarchical composition of SMNs forms an SMM. The metrics in SMNs can be defined in terms of logical operations. All nodes can be colored or left blank. The default coloring scheme of the MVS imitates traffic lights: red stands for insufficient level, yellow for intermediate level, and green for sufficient level [10].

An important challenge is that many objects of the SuI system architecture are *unmanaged* [11]: they are not within the Administration Domain (AD) of the stakeholder carrying out security management and/or measurement of the SuI. This is the case in many cloud services especially from the CSU perspective. Direct security measurements are not possible for an unmanaged object. However, a *trust value*, a certain value representing the amount of trust that the security of the object is adequate, can be associated with the object [12].

C. Utilization of SMM with Monitoring

As the SMM contains both prioritized risk information and detailed-level metrics related to the risks, it can be used to guide the monitoring activities. The SMM makes monitoring *meaningful* from the original security objectives perspective. The interaction between the SMM and monitoring should support also information flow to the opposite direction: in addition to actual measurement or monitoring results, it can potentially offer information of new types of vulnerabilities, risks and other information relevant to security effectiveness evaluation. This up-to-date feedback information should be incorporated to the SMM. Therefore, it is important to build a monitoring system in a way that it allows for interpretation of the results from a wider perspective.

IV. EXPERIMENTATION SYSTEM

In the following, we describe our experimentation set-up and experiences from a specific security configuration correctness measurement scenario in a private cloud service. The example is mainly related to R2.

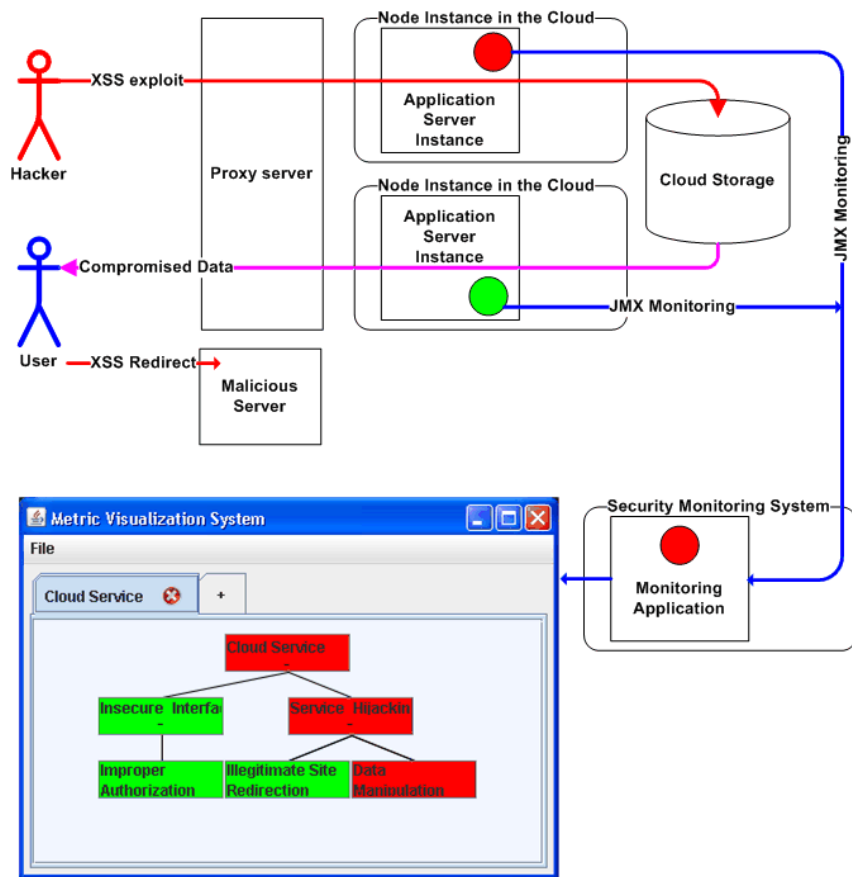


Figure 1. Hostile data injection and detection with a screenshot from the MVS tool.

A. Example Security Exploit

The following example illustrates a Cloud system with a XSS (Cross-Site-Scripting) weakness and a monitoring solution to detect the security issues. In this example, the Cloud API does not have sufficient data validation and a malicious user is able to inject a redirection Java script in to the Cloud storage. Other users utilizing the Cloud API may access the corrupted data assuming it is trusted data and the user interface will then redirect the victims to a malicious web site.

B. Monitoring Security Correctness

Let us assume that initially the SMM contains appropriate metrics pointing the details which security configuration parameters should be measured (monitored).

The monitoring system attempts to detect issues independently and separately from the application logic validation rules. In this example each of the application servers in the Cloud are bundled with remote management software components that are capable of accessing the live memory objects without interfering with the business logic. The remote monitoring application monitors all active application server instances in the cloud for any suspicious activities as defined in the SMM model shown in the MVS tool, see Figure 1. Once

the malicious transaction is detected, the MVS evaluates the failed tests and marks the corresponding tree structure in color red. The SMM should be updated to reflect the latest findings from the monitoring system. If there are already metrics matching to the findings, their measured values are updated. If not, the need for additional metrics should be considered.

C. Experimental Monitoring Set-up

For experimentation of this example, a testing environment was built on UEC (Ubuntu Eucalyptus Cloud) [13] containing multiple instances of Apache Tomcat [14] web application servers.

Our experimental security monitoring set-up utilizes a non-invasive technique called JMX (Java Management Extensions) [15] which allows remote monitoring of registered object resources. JMX is included in the Java Standard Edition [16]. This technology is limited to applications implemented in the Java programming language, but with some effort it is possible to add similar features to applications implemented in other languages as well. The remote monitoring capability from one centralized location is ideal for cloud environments so that the monitoring itself does not interfere the application execution.

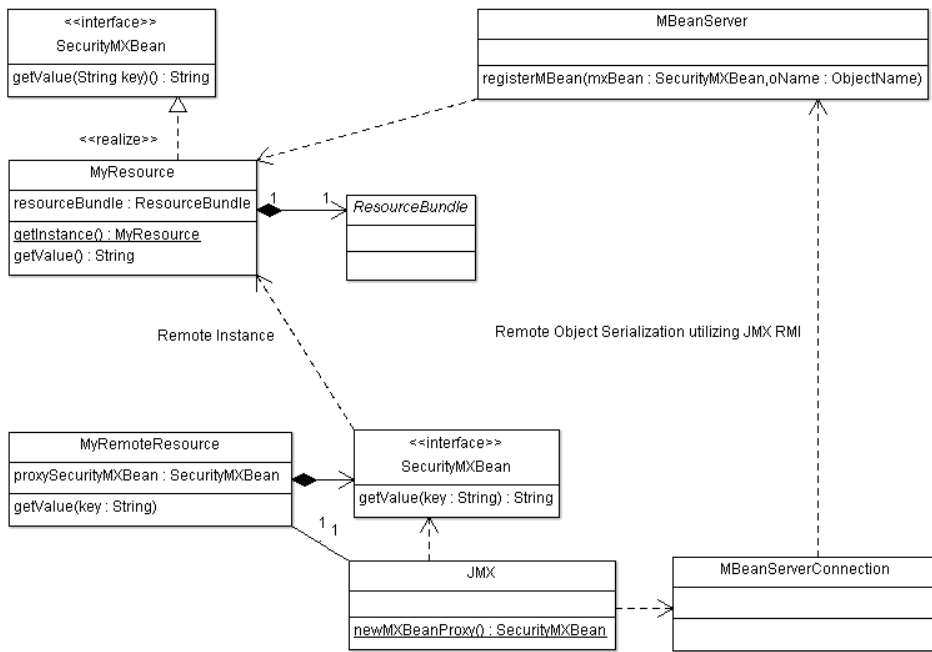


Figure 2. Class diagram of the Managed Bean monitoring system.

First we need to enable the Java Virtual Machine to allow remote management. The code block below shows how remote management is opened at port 1199 with no authentication and no SSL (Secure Sockets Layer) transport level security enabled:

```
CATALINA_OPTS="-Dcom.sun.management.jmxremote \
-Dcom.sun.management.jmxremote.port=1199 \
-Dcom.sun.management.jmxremote.ssl=false \
-Dcom.sun.management.jmxremote.authenticate=false"
```

Figure 2 shows a class diagram of the monitoring system. This example has a security sensitive singleton instance of class MyResource. MyResource contains sensitive application properties in its ResourceBundle instance.

The example shown in the class diagram demonstrates how method 'getValue' is exposed for read-only monitoring of a 'value' attribute from a remote location. In JMX, this can be achieved when the class implements a Platform MBean (Managed Bean) [17] interface and the MBean defines all the methods available remotely. On the observing side there is a need for a copy of the MBean interface in order to create remote clone objects. Java Standard Edition version 6 provides the JMX and MBeanServerConnection classes to handle the network connection, object serialization and the creation of the proxy object. The MBean needs to be registered to the server with a given object name. This object name will be the reference to the remote client code as shown in the code below. This registration should be done at the initialization phase of the application key resources. One way to do so is to implement your own ServletContextListener and register the MBeans in its contextInitialized method. This will guarantee that the code block gets processed before the web application

starts accepting any external requests. Otherwise, it is possible that the remote code will not be able to create a clone object when the references are not properly set.

```
SecurityMXBean mxBean = MyResource.getInstance();

MBeanServer mbs =
    ManagementFactory.getPlatformMBeanServer();

ObjectName objectName =
    new ObjectName("test:type=MyResource");

mbs.registerMBean(mxBean, objectName);
```

On the remote server, a proxy object instance of the MBean is created:

```
String jmxUrlString =
    "service:jmx:rmi:///jndi/rmi://"
    +serverName+": "+portNumber+"/jmxrmi";

JMXConnector connector =
    JMXConnectorFactory.connect(
        new JMXServiceURL(jmxUrlString));

MBeanServerConnection connection =
    connector.getMBeanServerConnection();

ObjectName objectName =
    new ObjectName("test:type=MyResource");

SecurityMXBean securityMXBeanProxy =
    JMX.newMXBeanProxy(connection, objectName,
        SecurityMXBean.class);

String value = securityMXBeanProxy.getValue(key);
```

This proxy object is a remote clone instance of the MXBean interface utilizing the Java RMI (Remote Method Invocation) technology. First we need to connect to a JMX remote server as discussed before. It should be noted that in cloud environment each node instance has a unique private IP (Internet Protocol) address. Some custom configuration may be required to access the node instances from the outside of the Node servers. Once the connection is established, it is known that the node instance is running and available, and monitoring of the registered MXBeans can be started. After this, the proxy object is created by referring to the same object name as it was defined in the code above. When the proxy instance is present, the available methods defined in the interface are called as a local object instance. The proxy object will call the remote `getValue` method from one of the cloud instance servers. If the return value of this remote method call differs from what the reference value is, there is a problem that needs to be analyzed more closely.

D. Experiences from Experimentation

According to our trials in the experimentation system, as shown in Figure 4, the monitoring approach discussed above is feasible for remote security monitoring in the Cloud. JMX technology is ideal for memory monitoring because memory values can be investigated in a non-intrusive and asynchronous way during run-time. For security monitoring, JMX technology offers flexibility of monitoring critical memory objects and other resources. It can be used to passive non-intrusive monitoring of different memory objects and analysis of them. Monitoring of different properties requires only minimal code addition to the software. Active monitoring can incorporate additional logging, prohibition of programme execution, or shutdown, triggered from a security alarm. However, a complex monitoring task can require remarkable code changes and can be intrusive to. The technical possibilities of the monitoring approach are constrained by the security policy of Java virtual machine. Moreover, access rights inherited from Operating System, such as file access and network input/output rights, set constraints to monitoring. Unmanaged establishment of a JMX connection can make the system vulnerable. The JMX port should be protected from unauthorized use using authentication and authorization mechanisms, VPN (Virtual Private Network) or SSL.

V. RELATED WORK

State-of-the-art security metrics approaches include security requirement decomposition [1], [18], attack surface metrics [19], security assurance metrics [20], security development lifecycle metrics [21], and standardization and recommendation efforts, such as Common Criteria [22]. Comprehensive overviews of security metrics approaches and objectives are found in, for example, [23–25].

Shao et al. introduce a runtime monitoring approach for the Cloud in [9], concentrating on QoS (Quality of Service) aspects. Their model, RCMC (Runtime Model for Cloud Monitoring) uses multiple monitoring techniques gather data from the cloud. However, security monitoring is not discussed in particular. Chazalet [26] discusses SLA (Service Level Agreements) compliance checking in cloud environments and

uses JMX technology in the prototype implementation. Their checking approach allows separating concerns related to the probes, information collection and monitoring and contract compliance checking. However, security monitoring is not discussed in particular in these contributions. The SLA monitoring approach in the European Commission's Framework 7 SLA@SOI project [27] relies on EVEREST+ [28] monitoring engine. De Chaves et al. [29] propose a monitoring approach called PCMONS for private clouds. The architecture of PCMONS includes node information gathering, data integration, virtual machine monitoring, configuration generation, monitoring tool server and a database approach. The metrics approach of the current paper can be integrated to PCMONS.

In addition, there are practical applications, where JMX technology is used to monitor CPU, thread and memory properties. It can be enhanced also for software update, network connection, and input/output resource availability monitoring.

Our specific monitoring example can be enhanced to address file integrity by utilizing the POR (Proof of Retrieval) protocol [30]. This protocol supports verification that large files are not deleted or modified prior to retrieval without downloading them.

VI. CONCLUSIONS AND FUTURE WORK

Metrics-driven security monitoring can potentially play a significant role in improving security, privacy and trust in cloud environments. We have discussed how security metrics can be used to set goals for security monitoring. Monitoring helps to identify security threats, vulnerabilities, and enables assurance of the correctness and adequateness of deployed security controls. We have introduced a novel remote security monitoring approach for the Cloud, which is used to security correctness monitoring of critical memory sections. The monitoring approach offers also feedback to security metrics management. The approach was investigated in an experimental private cloud service environment and was based on Java Management Extensions technology and enables connections to several servers at the same time, and memory values can be investigated in a non-intrusive and asynchronous way during run-time. Critical memory objects can be monitored in a flexible way. Our future work includes application of the monitoring approach to a real-world cloud application involving business-critical information management.

REFERENCES

- [1] R. Savola and H. Abie, Development of measurable security for a distributed messaging system, *Int. Journal on Advances in Security*, Vol. 2, No. 4, 2009, pp. 358–380.
- [2] R. Savola, "Towards a risk-driven methodology for privacy metrics development," *PSA '10*, 20-22 Aug., 2010.
- [3] Cloud Security Alliance, "Top threats to cloud computing", Version 1.0. Downloaded from: www.cloudsecurityalliance.org [Accessed June 30, 2012].

- [4] R. Savola, "A security metrics taxonomization model for software-intensive systems," *Journal of Information Processing Systems*, Vol. 5, No. 4, Dec. 2009, pp. 197–206.
- [5] W. Jansen, "Directions in security metrics research," U.S. National Institute of Standards and Technology, NISTIR 7564, Apr. 2009, 21 p.
- [6] Information Technology Security Evaluation Criteria (ITSEC), Version 1.2, Commission for the European Communities, 1991.
- [7] T. Kanstrén, R. Savola, A. Evesti, H. Pentikäinen, A. Hecker, M. Ouedraogo, K. Hätönen, P. Halonen, C. Blad, O. López and S. Ros, "Towards an abstraction layer for security assurance measurements (invited paper)," *Proceedings of ECSA: Companion Volume*, pp. 189–196.
- [8] T. Kanstrén, R. Savola, S. Haddad and A. Hecker, "An adaptive and dependable distributed monitoring framework," *Int. Journal on Advances in Security*, Vol. 4, Nos. 1 & 2, 2011, pp. 80–94.
- [9] J. Shao, H. Wei, Q. Wang and H. Mei, "A runtime model based monitoring approach for cloud," *IEEE 3rd Int. Conf. on Cloud Computing*, Miami, Florida, 5-10 July, 2010.
- [10] R. Savola and P. Heinonen, "A visualization and modeling tool for security metrics and measurements management," *Proceedings of ISSA 2011*, Johannesburg, South Africa, 8 p.
- [11] M. Ouedraogo, D. Khadraoui, B. de Rémont, E. Dubois, and H. Mouratidis, "Deployment of a security assurance monitoring framework for telecommunication service infrastructure on a VoIP system," *Proceedings of NTMS '98*.
- [12] R. Savola, H. Pentikäinen, and M. Ouedraogo, "Towards security effectiveness measurement utilizing risk-based security assurance," *Proceedings of ISSA 2010*, Aug. 2–4, 2010, Sandton, South Africa, 8 p.
- [13] D. Nurmi, R. Wolski, C. Grzegorzczak, G. Obertelli, S. Soman, L. Youseff, D. Zagorodnov, "The Eucalyptus open-source cloud-computing system," *CCGRID 2009*, pp. 124–131.
- [14] D. Chetty, "Tomcat 6 developer's guide," Packt Publishing, 2009.
- [15] Java Management Extensions (JMX) Technology. Oracle Sun Developer Network. Web site: <http://java.sun.com/javase/technologies/core/mntrmgmt/javamanagement/> [Accessed June 30, 2012].
- [16] J. Gosling, B. Joy, G. Steele, G. Bracha, "Java™ language specification," 3rd Edition, Addison-Wesley, 2005.
- [17] E. McManus, "What is an MXBean?" *Java.net*. Web site: http://weblogs.java.net/blog/emcmanus/archive/2006/02/what_is_an_mx_be.html [Accessed June 30, 2012].
- [18] C. Wang and W. A. Wulf, "Towards a framework for security measurement", 20th National Information Systems Security Conference, Baltimore, MD, Oct. 1997, pp. 522–533.
- [19] M. Howard, J. Pincus, and J. M. Wing, "Measuring relative attack surfaces", *Workshop on Advanced Developments in Software and Systems Security*, 2003.
- [20] E. Bulut, D. Khadraoui, and B. Marquet, "Multi-agent based security assurance monitoring system for telecommunication infrastructures," *CNIS 2007*, Berkeley, CA, USA, Sep. 24-27, 2007, 6 p.
- [21] S. Chandra and R. A. Khan, "Object oriented software security estimation life cycle – Design phase perspective", *Journal of Software Engineering*, 2008, Vol. 2, Issue 1, pp. 39-46.
- [22] ISO/IEC 15408-1:2005, "Common Criteria for information technology security evaluation – Part 1: Introduction and general model", ISO/IEC, 2005.
- [23] D. S. Hermann, "Complete guide to security and privacy metrics – measuring regulatory compliance, operational resilience and ROI," Auerbach Publications, 2007, 824 p.
- [24] A. Jaquith, "Security metrics: Replacing fear, uncertainty and doubt," Addison-Wesley, 2007.
- [25] N. Bartol, B. Bates, K.M. Goertzel and T. Winograd, "Measuring cyber security and information assurance: A state-of-the-art report," Information Assurance Technology Analysis Center (IATAC), May 2009.
- [26] A. Chazalet, "Service level checking in the cloud computing context," 3rd Int. Conf. on Cloud Computing, 2010, pp. 297–304.
- [27] P. Wieder, J.M. Butler, W. Theilmann, and R. Yahyapour (Eds.), "Service Level Agreements for cloud computing," Springer, 2011, 357 p.
- [28] D. Lorenzoli and G. Spanoudakis, "EVEREST+: Runtime SLA violations prediction," *Proc. of the 5th Middleware for Service-oriented Computing Workshop (MW4SOC '10)*, 2010, pp. 13–18.
- [29] S.A. de Chaves, R.B. Uriarte, and C.B. Westphall, "Toward an architecture for monitoring private clouds," *IEEE Communications*, 49 (12), 2011, pp. 130–137.
- [30] A. Juels and B.K. Kaliski, "PORs: Proofs of Retrievability for large files," *IACR Cryptology ePrint Archive*, ACM CCS, 2007, pp. 584–597.

Real Time Decision Support Systems for Mobile Users in Intelligent Cities

Alfio Costanzo¹, Alberto Faro¹

¹ *Department of Electrical, Electronics and Computer Engineering,
University of Catania, viale A. Doria 6,
Catania, 95125, Italy*

ABSTRACT - Current user mobility is supported by GPS navigation systems mainly based on average traffic conditions, whereas we know that both mobility and logistics operations could be facilitated if mobile users may choose the path to destination and the loading and unloading circuits using navigators that take into account the current traffic and weather conditions. Also, personal data, such as health status and age, should be taken into account by this new generation of navigators to suggest proper mobility to users. For these reasons, the paper aims at proposing a real time decision support system (DSS) that helps mobile users to reach their destination taking into account both the main external conditions and personal constraints. In the paper we propose that such DSS operates at a semantic layer on the top of the proprietary applications so that its suggestions take into account all the city datasets, while proprietary applications may continue to carry out implementation dependent functionalities, such as communicating with mobiles, taking traffic measures, launching specific alerts. The paper not only shows how structuring such a semantic layer, but also illustrates how providing such location based services to the user mobiles using JQMobile and Flash Builder frameworks.

I. INTRODUCTION

Both mobility and logistics operations could be facilitated if mobile users may choose the path to destination and the loading and unloading circuits taking into account the current traffic and weather conditions.

To achieve such ubiquitous intelligent mobility system, we have suggested in [1] to implement in the city four main layers: a) a communication network to exchange information with mobile users (e.g., ad hoc wifi net, GPRS, etc.), b) a sensing infrastructure to measure traffic by *in situ technologies*, by *floating car data* techniques or by *perceptions* issued by authorized people working on the territory, e.g., [2], [3], [4], c) an application layer where all the sensed data are collected in proprietary datasets and processed according to simple queries or suitable mathematical models to inform timely the users, and d) a semantic layer, where the proprietary datasets are transformed into standard datasets in RDF format [5] so that they can be integrated with other datasets to give rise to a distributed city Data Warehouse (DWH) consisting of many datasets resident on different machines that can be viewed as a single semantic information system.

This architectural choice allows the software at communication, monitoring and application layers to perform their functions according to proprietary solutions that meet implementation dependent management rules, e.g.,

privacy constraints, coding strategies to fully exploit the communication bandwidth, and especially signal and image processing algorithms and infrastructures to take traffic data in real time [6], [7] [8].

At the semantic layer all the datasets are integrated into an RDF real time DWH to be used by an intelligent software to help the mobile users in finding the most suitable service for their current needs, e.g. the best route to reach the park nearest to their destination or to reach the pharmacy currently open or to reach the nearest bank provided with working cash machine and so on.

Such Decision Support System (DSS) depends not only on the current traffic conditions but also on weather, and on personal data such as health status and age of the mobile users. In principle, the conditions to be considered by DSS could be stored only on the city information system, but a solution that is based on conditions partially stored on the city information system and partially on the user devices is also envisaged so that the personal information that influences DSS stays on the mobile under the direct user control.

For these reasons, the paper aims at proposing a real time DSS that helps mobile users to reach their destination taking into account both the main external conditions and personal constraints, and in which the user mobiles play an active role. Sect.2 illustrates how the semantic layer of a real time city DWH may be structured to provide the above mentioned DSS, whereas sect.3 shows two alternative solutions of how supporting mobile user decisions by a Fuzzy Logic framework [9]: a) one based on an Ruby on Rails (RoR) server [10] that manages all the rules of the fuzzy DSS, including the ones dealing with the user constraints, to suggest, using JQMobile scripts [11], the most useful destinations and related paths on the user mobile with a Google Maps based interface, and b) the second based on a solution where the RoR server manages the rules of the DSS in cooperation with the user mobile. Moreover, sect.3 briefly outlines how an user mobile provided with a Flash Builder [12] based application would play the sought active role in the future intelligent city scenarios envisaged by the latter solution.

II. BASIC INTELLIGENT CITY SERVICES

At the core of an effective DSS for mobile users is the knowledge of the time and distance to reach a sought destination and the weather conditions during the travel. To

compute this basic space-time information we may use the matrices D and T, being $d(i, j)$ and $t(i, j)$ respectively the distance and the travel time of the road segment from intersection i to intersection j . As known, the matrices D and T are the input of numerous procedural algorithms that are able to find the minimum path in terms of distance and in terms of travel time between any pair of intersections chosen by the mobile user, e.g., [13], [14].

However, in our implementation, the minimum path finder is written in Prolog [15] since we found that its time performance is similar to the one of the procedural algorithms, whereas its functions may be powered by simply adding business rules to the rules governing the search of the minimum path to destination. Another advantage of using Prolog is that it allows us to find the circuit that a logistics company should follow to collect (or to deliver) goods at a prefixed set of points starting from the company address and coming back, at the end, to the same initial address in a minimum time.

Also, any Prolog program operates naturally at the semantic layer, where DSS should stay according to the architecture proposed in [1]. Indeed, the matrices D and T, that are the input of the program, are represented by the following *facts* that don't depend on the structure of a particular database:

$$d_road(i_a, i_b, d_{ab})$$

$$t_road(i_a, i_b, t_{ab})$$

where i_a , and i_b , are intersections, and d_{ab} and t_{ab} are the distance and the travel time associated to the road from i_a , to i_b . Deriving such facts from tables, e.g., from the city road tables in fig.1, is certainly possible, but it requires a specific conversion effort, whereas they may be derived in a straightforward way from the urban RDF scheme in fig.2.

Road Segment		City A
Attribute	Instance	
Identifier	P01	
Name	Etnea: Bellini	
Intersection A	Via Etnea - Via Pacini	
Intersection B	Via Etnea - Via Umberto	
Description	The main road in the downtown	
Distance	200 meters	
Walking	2	
Driving	0	

Travel Time		
Attribute	Instance	
Identifier	P01	
Travel Time	150 seconds	

Proximity Table		City B
Attribute	Instance	
Identifier	P01	
Intersection A	I01	
Intersection B	I77	
L	200 meters	
T	150 seconds	
W	both ways	
D	no	

Figure 1 - Different ways of representing the road network in two cities.

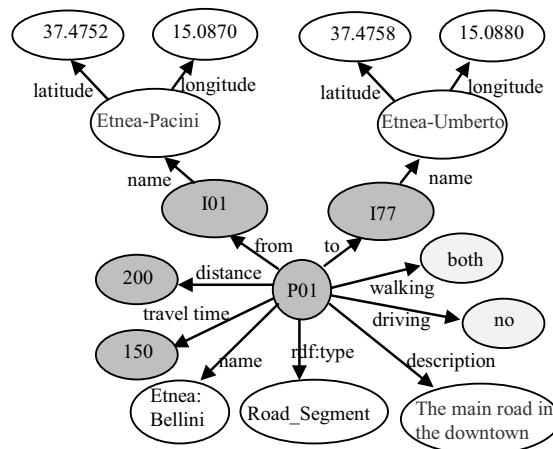


Figure 2 - RDF scheme of a road network valid for both cities A and B. The data in grey are used by the Prolog program.

In principle, a pure Prolog algorithm that computes the minimum path operates in a trivial way, i.e., first, it finds all the paths from the source to the destination, and then it determines the minimum one. Of course, this approach is not feasible in our case since the number of paths connecting two nodes of a traffic net is very great, even if the network is of medium-small complexity.

Thus, in our program we added an heuristics that stops the computation of a path if the time or distance to reach the current intersection from the source is greater than the distance or time of the path currently stored as the best one. This constraint, together with the one of avoiding to pass two times through the same road segment allows us to find the minimum path between two intersections in a very short time, e.g., about 50 milliseconds to find the best paths for the road network of the downtown of a medium city.

A comparison carried out at University between the time performance of the Prolog program and the one of a program developed using a procedural language has demonstrated that both the programs have similar time performance. The overall benchmark will be published in other works since it is outside the scope of the paper.

To appreciate the flexibility of such approach, let us note that the same Prolog program may be used, with a simple modification of the *facts*, to find the minimum path for other location services, e.g., a) to find the nearest park and b) the minimum path to reach the park nearest to the destination.

In the former case, it is enough to connect all the city parks to a fictitious node, let say *park_node*, with a road_segment featured by distance and travel time equal to zero, as shown in fig.3. Indeed, the minimum path to go from the intersection closest to the user position to *park_node* gives indirectly the minimum path to reach the nearest park, being the nearest park the one located at the penultimate node of such path. The same approach may be used to find the nearest hotel, pharmacy and so on.

In the latter case, first, we have to find the park at the minimum distance from the destination following the same approach described above, i.e., finding the minimum path, let say pt_1 , from the *park_node* to the destination and then

finding the minimum path from the current position to the park located at the second node of the previous path pt_1 . Also these two steps are illustrated in fig.3.

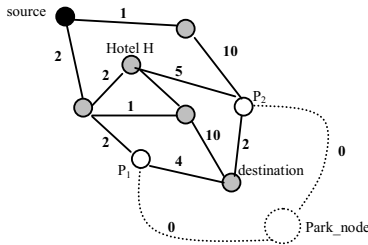


Figure 3. The nearest park is the penultimate node of the minimum path to reach the fictitious node denoted as Park_node, i.e., park P_1 . The park nearest to the destination is obtained by the second node of the minimum path connecting Park_node to the destination, i.e., park P_2 . The minimum path is the one that passes through the node of the Hotel H.

Having an estimation of the travel time to reach any destination better than the one provided by Google Maps allows us to extend the Location Based Services (LBSs) provided by Google Maps. In particular, we may suggest the destination services most suitable for the mobile user from both the distance and travel time points of view.

These real time LBSs may be further improved by using suitable fuzzy rules to provide location intelligence services so that "how much a service should be very near to the user" could be computed either in terms of distance or time by taking into account fuzzy rules expressed as follows:

"if condition 1 and and condition N are true then the service should be nearest to the user position".

As known, in fuzzy logic the truth degree of the conclusion is given by its membership degree to the fuzzy set associated to the conclusion itself, i.e., in our case the fuzzy set *nearest*, that, on its turn, is given by the minimum truth degree of the antecedents. This computation is very easy, since any fuzzy set may be represented by trapezoidal forms in the x-y setting, where the x-axis represents the variable that qualifies the set (e.g., the distance or the time in case of the fuzzy set *nearest*) and the y-axis represents the membership degree to the fuzzy set.

Thus, assuming that the fuzzy set *nearest service* is the one represented in fig.4, we have that if the membership degree of the antecedent is ≈ 1 , then the service should be between 0 and 300 meters, let say 150 meters, whereas if the degree \approx zero, then the service may be located at a greater distance, let say 250 meters. Of course, if the antecedent is equal to zero, the rule does not give any contribution to find the distance most suitable for the user.

However, in general, we have to meet more than one rule. As an example, typical fuzzy rules featuring location intelligence problems are as follows:

- if the weather is bad, then the service should be very near,
- if the current time is greater than 7 p.m. then the service may be located at medium distance
- if the user is an elderly people, the service should be very near,
- if the health status is not good the service should be very near.

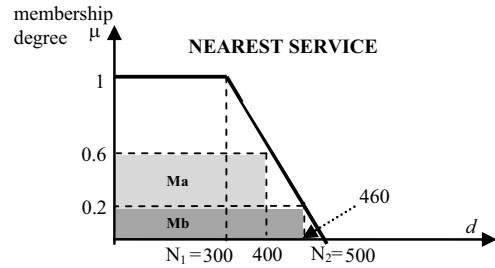


Figure 4. Fuzzy set *nearest service* depending on the distance d of the service from the current user position, e.g. if $d > 500$ meters the service cannot be considered as a *nearest service*, whereas if $d = 400$ meters it may be considered as near enough since $\mu = 0.6$.

As a consequence, we have to combine these fuzzy rules to obtain the service distance that best fit all the membership degrees associated to the weather, i.e., μ_w , day hour, i.e., μ_h , user age, i.e., μ_a , and health status, i.e., μ_s .

For sake of simplicity, we illustrate how obtaining the maximum distance acceptable by the user starting from the two fuzzy sets illustrated in fig.5 dealing with the weather and the user age, where we assume that the weather is evaluated from the raining degree, and the older age by the years of age. In particular, from a rain of medium intensity represented by 0.4 we obtain that the bad weather holds 0.6, whereas for an user of 45 years old we obtain that she/he may be considered a slightly elderly person, being $\mu = 0.2$.

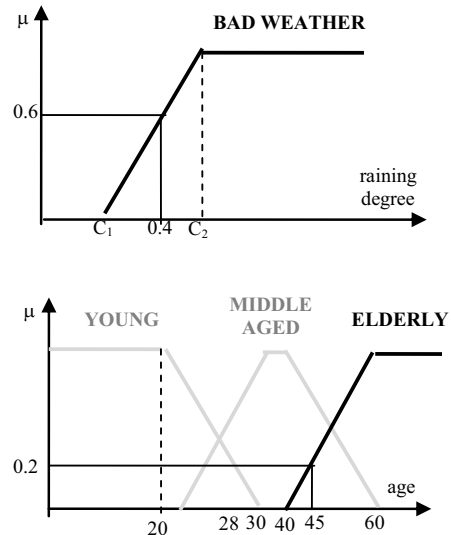


Figure 5. The fuzzy sets *bad weather* and *elderly*

The most simple rule to combine more than one antecedent is the one of considering the antecedent with the minimum μ . Since in the example the minimum is 0.2, we obtain that the maximum distance the user may accept to reach the destination from the current user position by intersecting the fuzzy set *nearest* with the horizontal line at $\mu = 0.2$, i.e., the maximum acceptable distance is between 0 and 460 meters, let say 230 meters. As a consequence, only the services that are at a distance less or equal to $d_{max} = 230$ meters will be suggested to the user by indicating also the best path to follow.

Let us note that:

- the method proposed may be extended to combine more than two the antecedents in a straightforward way.
- the above method may be followed also to compute the maximum time t_{max} the user may accept to reach the destination at condition of substituting in fig.4 the distance d by the time t , i.e. the time to reach a destination from the current user position, and putting $N_1 = 1$ minute, and $N_2 = 5$ minutes.
- to better take into account the antecedents of the fuzzy rule we may adopt other ways to combine the membership degrees of the antecedents, such as the one of considering that the d_{max} is not given by the x-coordinate of the barycentre of the "mass" M_b in fig.4 related to the antecedent with the minimum μ , but by the x-coordinate of the barycenter of the "masses" obtained by all the antecedents of the rule, e.g., the masses M_a and M_b in fig.4, with the convention that the value of a "mass" is given by its area.

III. ROR SERVER AND USER MOBILE INTERFACE

Currently, the location based services outlined in the paper, including the proposed extensions dealing with the location intelligence functionalities, are available on both PCs and mobiles. In particular, as outlined in the introduction, two versions have been developed: a) one for either PCs and mobiles based on RoR and JQMobile, and b) the second for mobiles based on RoR and Flash Builder.

In the former version, all the functions are carried out by the RoR server with an interface that, thanks to the JQMobile features, has the same format in both PC and cellular phone. In the second version, the cellular phones display the information on Google Maps without any intervention of the server. In fact, as pointed in sect.II, the server computes the minimum path to the destination chosen by the user, or to the one identified by using the mentioned city and user rules expressed in fuzzy format, but it does not build the map for the user. Indeed, it stores the path in an XML file that will be read by the user mobile to display it on the relevant Google Map.

Let us note that both the server and the Flash Builder mobile display the path in the Google Maps framework by imposing that the Google Maps API *getdirections* interconnects an initial point to a final one by passing through the intersections computed by the Prolog program that have to be considered as waypoints. In our interface, the nearest services in terms of distance are shown by icons located within a circle whose radius is computed by the fuzzy rules described in the previous section, where the fuzzy set used to defuzzify is the one denoted by *nearest distance*, whereas the nearest services in terms of time are shown within a circle whose radius is given by the most distant service which may be achieved in the time derived by the fuzzy set *shortest time*.

As an example, fig.6 shows the case in which there is only one parking nearest to the destination address located at center; the screen on the right displays the walking path computed by the Prolog program. It is generally the same of the computed by Google Maps. Another example is given

in fig.7, that shows a case in which there are two parks nearest to the current position of the mobile user located at center. The driving paths from the current position to the parks, shown on the left and on the right, clarify that, although the geo-coordinates of the park on the right are very close to the ones of the current mobile user position, both the parks can be reached at the same time. Indeed, they are at the same road distance and the traffic condition is the same in all the roads of the circle. Generally, the driving path computed by the Prolog program differs from the one computed by Google Maps due to the current traffic conditions.

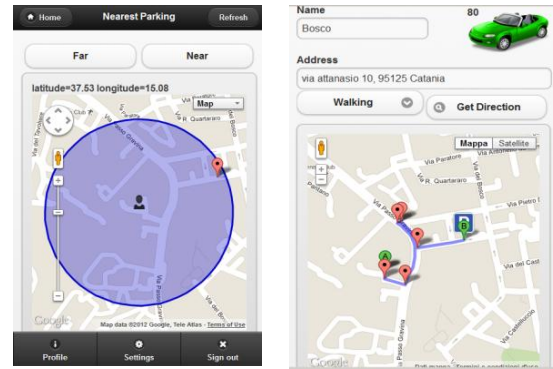


Figure 6. Only one parking is nearest to the destination located at center of the screen (left). The screen on the right displays the walking path.

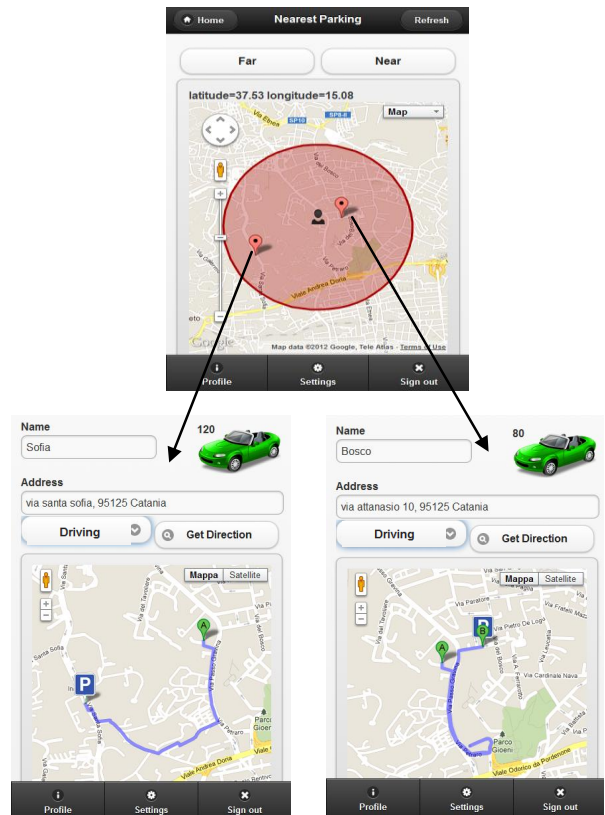


Figure 7. Two parks are very near to the current position of the mobile user located at center. The driving paths from the current position to the parks, shown on the left and on the right, clarify that, although the geo-coordinates of the park on the right are very close to the ones of the current mobile user position, both the parks can be reached at the same time.

Let us note that if the user is not provided with a mobile with a dedicated software abroad, then her/his personal data should be stored in the area of the server describing the user profile, whereas if the user is connected to the information system through a mobile provided with a Flash Builder software, then the personal data may be stored locally. To impose that the fuzzy DSS resident on the server will take into account the personal rules resident on the mobile, it is enough that the user issues her/his request for the nearest service together with current truth degree of each user constraint stored in her/his mobile. The use of an interface based on the familiar Google Maps guarantees that the mobile user may understand the information in a time that is compatible with the actions s/he has to take in the rapidly evolving use scenarios [16], [17].

IV. CONCLUDING REMARKS

The paper has shown how a Prolog algorithm powered by proper Fuzzy rules may be used to compute the minimum walking and driving paths depending on the traffic and weather conditions and on personal constraints. In general traffic and weather conditions are measured by a suitable sensing infrastructures, whereas personal data may be stored on either the server or on the user mobile. The suggested paths usually differ from the ones computed by Google Maps and are more effective to increase gas mileage of the cars, to decrease pollution in the city and to save time of the users.

The adoption of an urban RDF scheme favors the use of the Prolog program for supporting intelligent LBSs for the mobile users of a city. However, to avoid to map the RDF scheme of a city to the one of another one, it is envisaged the development of a common urban ontology [18], [19] holding for an entire region or country.

Also, a common ontology may favor the reuse of the Prolog software to develop similar use cases, e.g., [20] [21], similar aspects, e.g. [22], or similar software patterns, e.g., [23], [24].

REFERENCES

[1] Faro, A., Giordano, D. and Spampinato, C., Integrating Location Tracking, Traffic Monitoring and Semantics in a Layered ITS Architecture. *Intelligent Transport Systems, IET*, vol.5(3), 197-206, 2011

[2] Leduc, G., Road Traffic Data: Collection Methods and Applications, Working Papers on Energy, Transport and Climate Change, N.1, JRC European Commission, 47967 - 2008

[3] Faro, A., Giordano, D. and Spampinato, C., Evaluation of the Traffic Parameters in a Metropolitan Area by Fusing Visual Perceptions and CNN Processing of Webcam Images, *IEEE Transactions on Neural*

Networks, Vol.19 (6), 1108-1129, IEEE, 2008.

[4] Faro, A., Giordano, D. and Spampinato, C., Adaptive background modeling integrated with luminosity sensors and occlusion processing for reliable vehicle detection. *IEEE Transactions on Intelligent Transportation Systems*. Vol.12(4), 1398-1412, IEEE, 2011

[5] Powers, S. *Practical RDF*, O'Reilly Media, 2003

[6] Cannavò F., Nunnari G., Giordano D., Spampinato C., 2006. Variational Method for Image Denoising by Distributed Genetic Algorithms on GRID Environment. *Workshops on Enabling Technologies: Infrastructure for Collaborative Enterprises, WETICE '06*. 227-234, IEEE, 2006

[7] Faro, A., Giordano, D., Maiorana, F., Mining massive datasets by an unsupervised parallel clustering on a GRID: Novel algorithms and case study. *Future Generation Computer Systems*, Vol.27(6), 711-724, 2011

[8] Crisafi A., Giordano D., Spampinato C., GRILAB 1.0: Grid Image Processing Laboratory for Distributed Machine Vision Applications. . Proc. 17th IEEE Int on Enabling Technologies: Infrastructure for Collaborative Enterprises, WETICE '08, 188-191, IEEE, 2008

[9] Wang P.P., 2001. *Computing with words*. Wiley Interscience, 2001

[10] Hartl M., *Ruby on Rails 3*, Addison Wesley, 2011

[11] David M., *Developing Websites with jQuery Mobile*, Focal Press, 2011

[12] Corlan M., *Adobe Flash Platform Tooling: Flash Builder*, Adobe, 2009

[13] Zhan, F. B., Noon, C. E., Shortest Path Algorithms: An Evaluation Using Real Road Networks, *Transportation Science*, Vol.32(1), 65-73, 2008

[14] Sneyers J., et al., Dijkstra's algorithm with Fibonacci heaps: An executable description, Proc. 20th ACM Workshop on Logic Programming (WLP'06), ACM Press, 2006

[15] Bratko, I., *PROLOG Programming for Artificial Intelligence*, Addison-Wesley Educational Publishers Inc, 2011

[16] Giordano, D. Evolution of interactive graphical representations into a design language: a distributed cognition account, *International Journal of Human-Computer Studies*, Vol. 57(4), 317-345, 2002

[17] Faro, A., Giordano, D., Ontology, esthetics and creativity at the crossroads in Information System design, *Knowledge Based Systems*, vol.13 (7), 515-525, Elsevier, 2000

[18] Zhai, J., Jiang, J., Yu, Y. and Li, J.: Ontology-based Integrated Information Platform for Digital City, *IEEE Proc. of Wireless Communications, Networking and Mobile Comp.*, WiCOM '08, 2008.

[19] Faro, A., Giordano D., Musarra, A. 2003: Ontology Based Mobility Information System. Proc. of IEEE Systems, Men and Cybernetics , vol.5, 4288-4293, Washington, IEEE, 2003

[20] Mansar S.L., Marir F., Reijers H.A., Case-Based Reasoning as a Technique for Knowledge Management in Business Process Redesign, *Electronic Journal on Knowledge Management*, Volume 1 Issue 2, 2003 113-124 Academic Conferences Limited 2003

[21] Faro, A., Giordano, D. StoryNet: an Evolving Network of Cases to Learn Information Systems Design. *IEE Proceedings SOFTWARE*, vol.145(4), 119-127, 1998

[22] Jacobson I., Pan-Wei Ng, Aspect-oriented software development with use cases, Addison Wesley, 2004

[23] Faro, A., Giordano, D. Concept Formation from Design Cases: Why Reusing Experience and Why Not. *Knowledge Based Systems Journal*, vol.11(7), 437-448 , Elsevier, 1998

[24] Faro, A., Giordano, D., Design memories as evolutionary systems: socio-technical architecture and genetics, *IEEE Proc on Systems Man and Cybernetics*, vol.5, 4334-4339, IEEE, 2003

Lip reading using fuzzy logic network with memory

Stefan Badura

Faculty of management science and informatics
University of Zilina, Itall s.r.o.
Zilina, Slovakia
baduras@itall.sk

Martin Klimo, Ondrej Skvarek

Faculty of management science and informatics
University of Zilina
Zilina, Slovakia
{martin.klimo, ondrej.skvarek}@fri.uniza.sk

Abstract— This paper proposes a new approach for lip reading. Most existing systems for lip reading utilize a kind of neural network or hidden Markov models as classifiers. We propose a new approach where fuzzy combinational networks with fuzzy flip-flop memories are combined into one network. Our model introduces a hierarchical structure of this network, where single layers are contextually dependent. Experiments with fuzzy flip-flop network propose a new approach in the process of automatic lip-reading system where time dependence in inputs series is modeled with memories. Such approach provides possibilities for continuous speech recognition.

Keywords - lip reading, fuzzy logic, flip-flop, neural network, memory

I. INTRODUCTION

Lip reading is a process where a person or computer tries to recognize speech from visual information by watching speaker's lip movements. Information obtained from lip reading can enhance intelligibility of speech utterances under difficult listening conditions. Audio-Visual Automatic Speech Recognition (AVASR) systems use visual information to enhance Automatic Speech Recognition (ASR) systems [1]. The process of automatic lip-reading can be generally separated into two logical parts:

1. Feature extraction from face region, lip movements, etc.
2. Speech recognition - in most cases this part can be considered as a problem of pattern classification, where various classifiers can be used.

The goal of this paper is to introduce a new approach for the second part - recognition part. We propose a novel network model, where we combine fuzzy logical structures defined in [19] with fuzzy flip-flop described in [12] for lip reading. Our goal is to interconnect simple two class classifiers into robust network that can be considered as universal multiclass classifier for dynamic data. Hierarchical organization and layered structure introduces contextual modeled system, which is not difficult to understand and which provides suitable abilities for general task of speech recognition. Speech is non-stationary process; our goal is to model this property using mentioned memories and fuzzy combinational structures. In this paper a basic model of memories is used. We suppose that using memories and layered structure provide support for building a multi-classifier from simple classifiers, which is not always a simple task. Network is designed to be able to recognize continuous speech.

Many approaches exist for automatic lip-reading. Most of them use artificial neural networks (ANN) or hidden Markov models (HMM). We introduce a network structure that is based on fuzzy logic circuits and combinational structures. Proposed model is tested on vowels of Slovak language.

From existing methods used for the task of lip-reading, recurrent neural networks are often used. In [2] authors recognize silence and vowels, where an Elman topology of ANN is utilized and that is constructed from 3 layers. In [3],[4],[6] a time delay NN (TDNN) is used. In [5] a modified TDNN is introduced for the same purpose. Many researchers resorted to the hidden Markov model (HMM) since it performs well in audio speech recognition [7],[8],[9]. In [10] the Ergodic HMM is used as recognizer. In [11] is presented another system using a fuzzy logic. Our network design models time dependence in speech but in different way – with using memories.

Our network design exploits fuzzy logic structures and fuzzy flip-flops memories to create the fuzzy logic network with memory. Using fuzzy logic provides ability to work in continuous interval $\langle 0, 1 \rangle$, where various operations can be used. For modeling time dependence of inputs a memory is used. The basic idea is to maintain output on higher value (remember it) when it is recognize by combinational circuit and for this purpose the flip-flop circuit serves. The fuzzy flip-flops are defined in [12] and they find inspiration in binary logic circuits, where logic functions are replaced with fuzzy logic operation. From experiments in [12] some positive properties of fuzzy flip-flops are shown. According [13] as many different logic operations exist, then also many ways of definition of fuzzy flip-flops exist with various behavior. In [14],[15],[16] is shown, that fuzzy systems constructed from fuzzy flip-flops can be considered as universal approximators. In [17] the basic concept of fuzzy flip-flop is introduced, where conjunctive and disjunctive norm form is used for the RS flip-flop. In [18] can be a comparative study between fuzzy networks and ANN found. In [13] a J-K and D flip-flop are used as basic elements in ANN design. From existing works it is clear that fuzzy logic and fuzzy networks or networks built from basic flip-flop elements provide capabilities these can be used in classification problem.

This paper is organized as follows. Next section introduces proposed network design. In 3th section inputs for recognition process are briefly described. The 4th section discusses experimental results and the last section concludes this paper.

II. NETWORK DESIGN

The main goal of designed network is to verify its possibilities on the task of speech recognition - lip reading. Especially we concentrate on experiments with memories, because we suppose they provide well behavior for modeling dynamic properties of speech.

A. Structure

The topology consists of 2 layers, where each layer has a purpose see figure 1.

- *The 1st layer* - this layer tries to indicate simple properties. In the task of lip reading one property could be e.g. mouth position (open-close position).
- *The 2nd layer* - this layer tries to consider time dependence of detected properties.

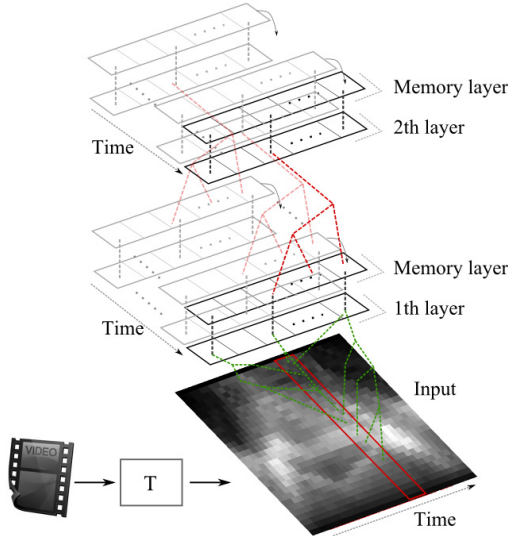


Figure 1. general network model. Layer 1 and 2 consist of trained structures.
The block T presents a image transformation into feature space.

Both layers consist of trained fuzzy combinational structures. Each structure is trained on one property against other properties. The training process, objective function and used fuzzy logic for one structure is defined and described in [19], where the best results are obtained using Lukasiewicz logic (see (1) and (2)) and objective function that maximizes distance between means of outputs classes obtained from structures for all inputs [19]. The equation (1) presents Lukasiewicz implication and in (2) is Lukasiewicz non, where y is an output and a, b are input parameters. The purpose of using such objective function is to maximize result to the value close to 1 for positive pattern and otherwise (see figure 2). Author in [19] shows that combinational structure built from described logic provides suitable capabilities for effective class recognition for that it was trained. Then we use this ability in our network design.

$$y = \min(1 - a + b, 1) \quad (1)$$

$$y = 1 - a \quad (2)$$

As it was already mentioned, the first layer tries to indicate some properties. In our case we define property as a number of a cluster. A set of training samples is grouped with Ward method into several clusters. Then each cluster represents one property. The number of groups was chosen as 15 (see the dendrogram at Figure 3 for given input data) where a slice for 15 groups is shown.

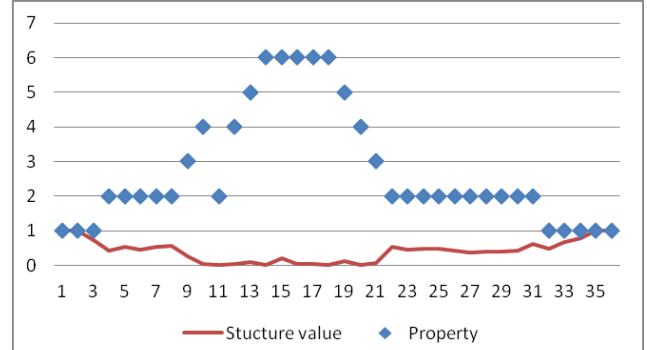


Figure 2. Dependence of structure output from proposed property sequence in the input. The output correspond to structure trained for 1th property.

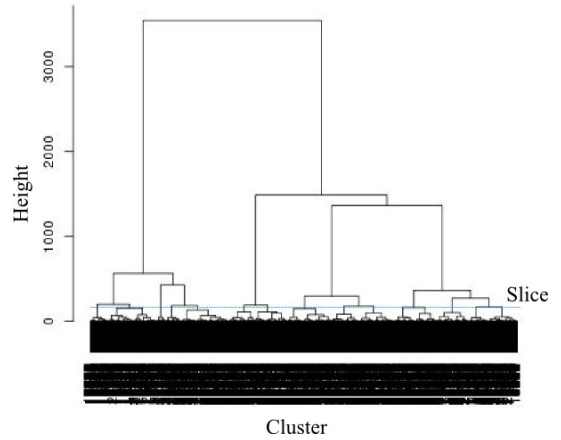


Figure 3. Dendrogram for input data.

The 1th layer consists of 15 structures, where each structure is trained for one property (one data cluster). Structures for second layer are trained for the output values of the first layer. We are using 23 vowels in our experiments, so the 2th layer is generated from structures trained on output sequences from the first layer for each vowel. The topology of the second layer is more complicated than the first layer, but the training principle of second layer is very similar to the first one. Difference is in objective function, where it is evaluated after time period for one vowel, when training structures.

B. Memory

Memory is important part, when time dependence of input data is modeled. In [12] a fuzzy flip-flop that provides abilities for sequential remembering of input signal was presented. If signal is close to the value 1, the flip-flop can remember its value – it is excited. If higher signal is proposed for longer time period, output stays excited also for longer time. In our experiments we used basic flip-flop as it is shown in Figure 4.

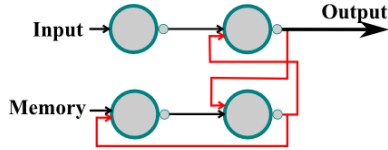


Figure 4. Basic flip-flop used as memory designed from NAND fuzzy logic operation.

Above each of introduced two layers could be a memory layer created from fuzzy flip-flops used. As it was described, the principle of memory is as follows: if signal is enough high (close to 1) for longer time period, memory supports it and otherwise. Using memory serves also for modeling time dependence. From other point of view on memory, it can be said, that memory enlarges the gap between stronger a weaker signals. In our experiments we use basic flip-flop memory as the figure 4 shows. The main goal of using memory is to provide ability for continuous speech recognition. We are interesting what the state of the network is after some time period. Structures these are fired on higher level are probably indicating input for which they were trained. This is the main idea of using memories. In this experiment we use memory just above the second layer.

III. LIP READING DATA

As inputs for the lip reading task, data of vowels extracted from video sequences are used. The flow chart on the image figure 5 shows the process of feature extraction. The figure 6 shows an example of extracted feature for video sequence. Median sieves are used for scale, space invariant feature extraction. Each column from the image on the figure 6 represents feature vector extracted from one video frame.

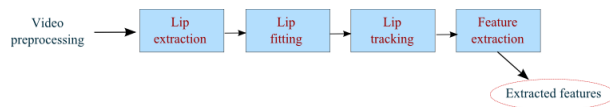


Figure 5. Flow chart for feature extraction process.

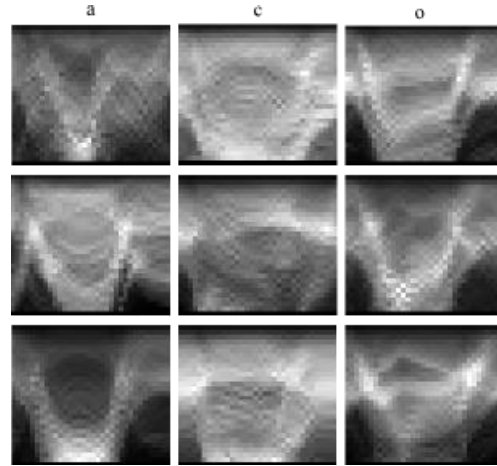


Figure 6. Example of extracted feature for lip-reading.

IV. EXPERIMENTS

For experiments a database of Slovak vowels was collected. Together 23 different vowels were recognized and each vowel was recorded 54 times. Next the whole dataset was splitted into training and testing subsets with the ratio 6:4. All structures of first and second layer were trained just on first subset. For the first layer Ward clustering algorithm was used for defining properties. This method separated time rows for each feature vector into 15 groups, these represent interesting properties. For evaluation of vowel recognition two different objective functions were used:

- U1 cumulative objective function over time
- U2 takes the value in the last time

When using U2 also another parameter was examined, and it is the fall time (θ value is proposed to network as input). In the figure 7 the best achieved results for both objective functions are shown. The results for U2 are interesting from sequential firing of structures, we are interested in the value at the end of recognized sequence – the most excited structure after some period time. The table 1 shows results obtained from experiments dependent on fall time and memory value for two mentioned objective functions.

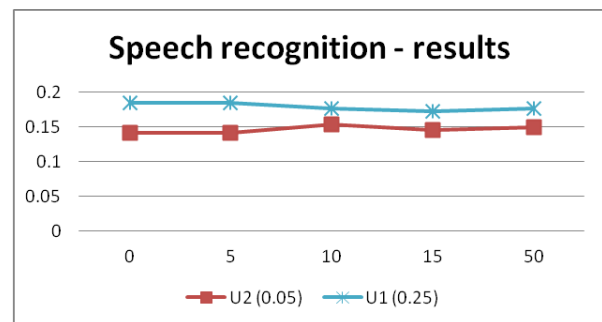


Figure 7. The best achieved results for both objective functions. The curve shows dependence of classification percentage from fall time.

TABLE I. EXPERIMENTAL RESULTS - POSITIVE RECOGNITION RATE IN %.
 ROWS PRESENTS RESULTS DEPENDING FROM FALL TIME AND COLUMNS
 PROPOSE RESULTS BASE FROM MEMORY VALUE AND OBJECTIVE FUNCTION.

Memory	0.05		0.15		0.25	
Fall time	U1	U2	U1	U2	U1	U2
0	16.41	14.16	15.84	12.90	18.51	12.34
5	16.41	14.16	16.54	14.58	18.51	13.74
10	16.12	15.28	16.83	14.02	17.67	12.34
15	16.12	14.58	17.11	14.44	17.25	10.93
50	16.12	15.00	16.97	6.45	17.67	4.347

The best results were obtained when using cumulative objective function over time. But from our experiments the results are interesting where memories were used, because of considering the state of the system in given time. In our experiments we tried to set memory and fall time value to be the most accurate and to get close with recognition rate to values that was obtained just with using combinational networks (no memory or fall time). The recognition rate of such network varied around 19.5% for U1. The table 1 shows results for training data. The best result for testing data moved around 10-12%, when using memories, or around 15% without memories. Compared to other methods used in our dataset the results somewhere in the middle. With simple KNN we obtained results around 30% of positive recognition and with Elman topology of recurrent neural network the results moved around 10-11%.

The difference between results for training and testing data set can be caused by small data set during training phase, especially for the second layer of our model.

V. CONCLUSION

In this paper novel approach for speech recognition has been introduced. It shows good behavior for the task of lip-reading. The final result obtained from experiments moved around 18-19%, which is considered as satisfied, but for effective speech recognition in the future it must be higher.

Interesting results are obtained when using memories, from experiments it is proved, that using memories leads to modeling the dynamic properties in input data. More complex comparison with other systems is not trivial task because of used database, language and vowels. Next experiments should cope with memory modeling. Memory in described experiment does not consider time occurrence of excitation. Other experiments those can be executed can concentrate on objective function for training especially of second layer. Important task that should be considered is diversity of used trained structures, because from other point of view proposed network design can be considered as multi classifier.

ACKNOWLEDGMENT

This work was supported by the Slovak Research and Development Agency under the contract No. VMSP-II-09.

REFERENCES

- [1] Scanlon, P., Reilly, R. Feature Analysis for Automatic Speechreading. In Proc. Int'l Workshop Multimedia Signal Processing, pp: 625–630, 2001.
- [2] Williams, R.J., Zipser, D.: A learning algorithm for continually running fully recurrent neural networks. In: Neural Computation. 1989.
- [3] Duchnowski, P., Hunke, M., Busching, D., Meier, U., Waibel, A.: Toward movement-invariant automatic lip-reading and speech recognition, 1995.
- [4] Duchnowski, P. – Meier, U. – Waibel, A.: See me, hear me: Integrating automatic speech recognition and lipreading. In: Proceedings of the ICSLP. 1994.
- [5] Prasad, V. K. – Stork, G. S. – Wolff, G. J.: Preprocessing video images for neural learning of lip reading. 1993.
- [6] Bregler, C., Omohundro, S. "Nonlinear manifold learning for visual speech recognition," in Proc. IEEE ICCV, pp. 494-499, 1995.
- [7] Goldschen, A.J. Continuous automatic speech recognition by lipreading, Ph.D. dissertation, George Washington Univ., Washington, DC, Sept. 1993.
- [8] Potamianos, G., Cosatto, E., Graf, H.P., Roe, D.B. Speaker independent audiovisual database for bimodal ASR. In Proc. European Tutorial Workshop Audiovisual Speech Processing, Rhodes, Greece, Sept. 1997.
- [9] Potamianos, G., Verma, A., Neti, C., Iyengar, G., Basu, S. A cascade image transform for speaker independent automatic speechreading. In Proc. IEEE Int. Conf. Multimedia, New York, Aug. 2000.
- [10] Sujatha, B., Santhaman, T. A Novel Approach Integrating geometric and Gabor Wavelet Approaches to Improve Visual Lipreading. Int. Journal of Soft Computing 5(1), 2010.
- [11] Silsbee P., Bovik, A. Computer lipreading for improved accuracy in automatic speech recognition, IEEE Trans. Speech Audio Processing, pp. 337-351, 1996.
- [12] Klímo, M., Boron, J. Dynamické vlastnosti pravdepodobných fuzzy klopných obvodov. In: proc. of ITAT (Informačné technológie – aplikácie a teória). 2009.
- [13] Lovassy, R., Kóczy, L. T., Gál, L. Analyzing Fuzzy Flip-Flops Based on Various Fuzzy Operations. In: Acta Technica Jaurinensis, Series Intelligentia Computatorica, vol.1, no 3, 2008.
- [14] Dubois, D., Grabisch, M., Prade, H. Gradual rules and the approximation of functions. In: Proc. of the 2nd International Fuzzy Systems Association Congress Iizuka. 1992.
- [15] Kosko, B.: Fuzzy Systems are Universal Approximators. In: Proc. Of the IEEE International Conference on Fuzzy Systems. 1992.
- [16] Wang, L. X.: Fuzzy Systems are Universal Approximators. In: Proc. Of the IEEE International Conference on Fuzzy Systems. 1992.
- [17] Hirota, K., Ozawa, K. Concept of fuzzy flip-flop. In: Preprints of 2nd IFSA Congress. 1987.
- [18] Lovassy, R., Kóczy, L. T., Gál, L. Robustness of Fuzzy Flip-Flop Based Neural Networks. In: CINTI 2010, 11th IEEE International Symposium on Computational Intelligence and Informatics. 2010.
- [19] Foltan, S. Speech recognition by means of fuzzy logical circuits. In 18th International Conference on Soft Computing, 2012.

Analysis of a Novel Audio Hash Function Based upon Stationary Wavelet Transform

Mahdi Nouri¹, Zahra Zeinolabedini²

Department of Electrical Engineering
Ghiasodin Institute of Higher Education
Abeyk, Iran

mnouri@mtu.edu¹, z.zeinolabedini@gmail.com²

Nooshin Farhangian³, Nasim Fekri⁴

Department of Electrical Engineering
Ghiasodin Institute of Higher Education
Abeyk, Iran

n.farhangian@gmail.com³, n67.fekri@gmail.com⁴

Abstract— Robust hashing for multimedia authentication is an emerging research area. Audio hash functions provide a tool for fast and reliable identification of content. A different key-dependent robust audio hashing based upon speech construction model is proposed in this article. The proposed audio hash function is based on the essential frequency series. Robust hash is calculated based on linear spectrum frequencies (LSFs) which model the verbal territory. The correlation between LSFs is decoupled by Stationary wavelet transform (SWT). A randomization structure controlled by a secret key is used in hash generation for random feature selection. The audio hash function is key-dependent and collision resistant. Temporarily, it is extremely robust to content protective operations besides having high accuracy of tampering localization. They are found, the first, to perform very adequately in identification and verification tests, and the second, to be very robust to a large range of attacks. Furthermore, it can be addressed the issue of security of hashes and proposed a keying technique, and thereby a key-dependent audio hash function.

Keywords—components; SWT; content-based authentication; robust hashing; least-square periodicity

I. INTRODUCTION

A perceptual audio hash function offers a tool for firm and reliable identification of content. A new audio hash functions is presented based upon précis of the time-frequency spectral characteristics of an audio file. In this work, the procedure is developed for summarizing a stretched audio signal into a short signature sequence, which can then be utilized to recognize the original record. The main encourage is obtaining audio hash functions which are insensitive to “reasonable” signal processing and edition process, for example compression, filtering, conversion of sampling rate and so forth, thus so sensitive to any changes in content. Perceptual audio hash functions can be used as an instrument to search for an exact record in a databank, to sense content tampering attacks, and so forth [1]. This kind of audio hash functions map speech to a short binary string based on the speech’s perceptual properties, is proposed as a different solution for automatic speech keying and speech content authentication. Contrasting to the traditional cryptographic audio hash function, which is exceedingly sensitive to the input data, the audio hash function permits for some changes

of the audio file whereas distinguishing all speech parts from another. Generally, the audio hash function requires two properties: discrimination, which means that perceptually dissimilar audio parts must have different hash vectors, and perceptual robustness, which means that identical speech parts should have the similar hash vector.

Because of wide application in automatic speech keying and speech authentication, the perceptual audio hash function has been widely studied recently [2-5]. However, there are few audio hash functions available. A compressed domain audio hash structure participated with a mixed excitation linear prediction codec is proposed in [6]. It applies partial bits of the speech bit stream with the linear spectral frequencies; the hash vector round based upon NMF of linear prediction coefficients is proposed [7]. The two features of the perceptual audio hash function which are so imperative, uniqueness and robustness. The uniqueness condition which is sometimes called randomness, infers that the hash sequence should reveal the content of the audio file in a unique manner.

There exist a number of audio hashing procedures in the literature. Mihcak and Venkatesan [9] extract statistical factors from arbitrarily particular regions of the time-frequency demonstration of the signal; Haitsma et al. offered an audio hashing algorithm [8], where the hash extraction system is based upon measuring the threshold of the energy changes between frequency bands, in another algorithm. In this work, one perceptual audio hashing procedure is investigated that operate in the frequency domain, and use the inherent periodicity of audio signals.

Wavelets are mathematical functions that cut up data into different frequency components, and then study each component with a resolution matched to its scale. They have advantages over traditional Fourier methods in analyzing physical situations where the signal contains discontinuities and sharp spikes. Wavelets were developed independently in the fields of mathematics, quantum physics, electrical engineering, and seismic geology. Interchanges between these fields during the last ten years have led to many new wavelet applications such as image compression, turbulence, human vision, radar, and earthquake prediction. This paper introduces

wavelets to the interested technical person outside of the digital signal processing field. The history of wavelets beginning can be described with Fourier, compare wavelet transforms with Fourier transforms, state properties and other special aspects of wavelets, and finish with some interesting applications such as image compression, musical tones, and de-noising noisy data.

In this paper, the stationary wavelet transform (SWT) [11] is a wavelet transform algorithm designed to overcome the lack of translation-invariance of the discrete wavelet transform (DWT). Translation-invariance is achieved by removing the downsamplers and upsamplers in the DWT and upsampling the filter coefficients by a factor of 2^{j-1} in the j -th level of the algorithm [12]. The SWT is an inherently redundant scheme as the output of each level of SWT contains the same number of samples as the input, so for a decomposition of N levels there is a redundancy of N in the wavelet coefficients. This algorithm is more famously known as "*algorithme à trous*" in French which refers to inserting zeros in the filters.

II. PERIODICITY-BASED HASH FUNCTIONS

A. Periodicity measure by a correlation-based analysis

The first peak of the autocorrelation of the linear prediction residue indicates the pitch period and is commonly used as a pitch estimator. This correlation-based periodicity estimate, called CPE, has the following expression:

$$\widehat{P}_0 = \begin{cases} \arg \max R(K), \text{ for } K \neq 0 \text{ if } R(\widehat{P}_0) \geq 0.5 \\ 0 \text{ if } R(\widehat{P}_0) < 0.5 \end{cases} \quad (1)$$

The efficacy of the CPE method is enhanced by a four-tap prediction and decimation process. The advantage of the correlation-based method is that it requires about three times less computation as compared to the parametric estimation.

B. The simulated attacks

The template is used to format your paper and style the text. All margins, column widths, line spaces, and text fonts are prescribed; please do not alter them. You may note peculiarities. For example, the head margin in this template measures proportionately more than is customary. This measurement and others are deliberate, using specifications that anticipate your paper as one part of the entire proceedings, and not as an independent document. Please do not revise any of the current designations.

In this section, we focus on transform-domain hash functions in contrast to the previous section, where we essentially worked on the time domain. More specifically, the audio signal is divided into possibly overlapping frames and each frame is represented by its mel-frequency cepstral coefficients (MFCCs), which are short-term spectral-based features [15]. Singular value decomposition (SVD) further summarizes these features. Note that in the SVD-based method we use the original signal and not its low pass filtered version, as in the periodicity-based schemes.

$$R = \left| \frac{N \sum_f X(f)Y(f) - \sum_f X(f) \sum_f Y(f)}{\sqrt{[N \sum_f X^2(f) - (\sum_f X(f))^2][N \sum_f Y^2(f) - (\sum_f Y(f))^2]}} \right| \quad (2)$$

$$d = \frac{1}{N} \sqrt{\sum_f^N (X(f) - Y(f))^2} \quad (3)$$

III. PROPOSED ROBUST SPEECH HASHING

The proposed scheme has two phases: feature extraction and hash modeling. In the feature extraction phase, perceptual features are extracted from audio signals such that the hash values could be robust to content-preserving operations. In the following hash modeling phase, a random secret key is employed to generate the secure hash sequence. Here, the extracted feature vectors are compressed to be compact and coded with the secret key to be randomized. The whole process is detailed in the following subsections.

A. Feature Extraction

Linear prediction coding (LPC) is widely used for speech coding and recognition, in which voice is modeled as the response of a vocal tract filter to a glottal excitation [9]. LPC coefficients model the slowly varying transfer function of the vocal tract. The varying transfer function determines the vowels which are important to speech perception. As a linear spectral representation of LPC coefficients, linear spectral frequencies (LSFs) have been widely used in speech coding and other speech processing applications. Therefore, as the content-based robust speech feature, LSFs are employed to generate the hash value in this letter.

To verify the discriminative and robust nature of the proposed audio hash function, it was applied to 1,000 audio clips (16 bits signed, 8 kHz) with various contents. On the waveform, the X-axis represents time, with past to future moving from left to right. The Y-axis represents intensity of the sound. The waveform exactly reflects the nature of a sound, which is just a series of fluctuations across time. Note that the signal is essentially periodic and repeating, though it has a somewhat random character. In fact, most sounds can be simply thought of as a combination of different repeating signals with various amplitudes and frequencies. Essentially, if an individual spectrum is thought of as a cross-section of a mountain, for example, then a spectrogram corresponds to a topographical map, composed of many spectra arranged side by side on their ends. A graph of a spectrum is produced by taking a small area around a single point and determining the amplitude and frequency of the signals immediately surrounding that point. They are then plotted on to the graph above, where the X-axis is frequency in Hz, and the Y-axis is amplitude in decibels.

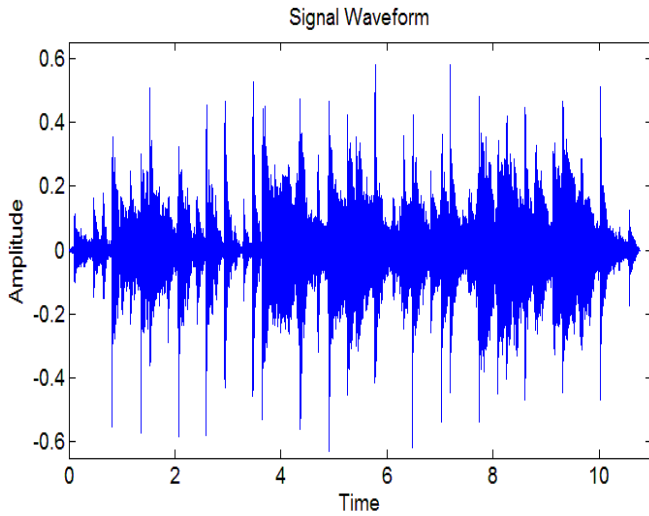


Figure 1. Intensity of the sound per time of main audio file

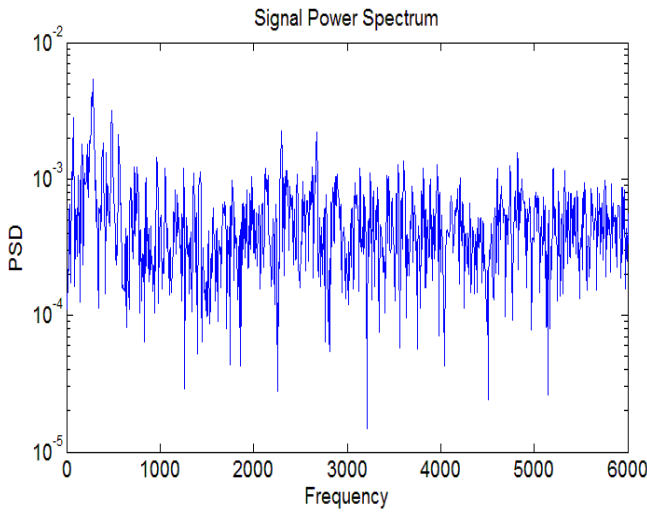


Figure 2. Power spectrum density per frequency of main audio file

Now it is easy to determine the frequency of waves of any amplitude, and vice versa, which would have been very difficult to determine from the waveform above. Bear in mind, however, that this spectrum is only of a single point, meaning that an entire speech sound cannot be easily analyzed using a spectrum.

Therefore, a random matrix of SWT coefficients is obtained. There are many alternative transformations available, such as DCT, FFT and PCA. Because of the good energy compaction property, SWT is used to decorrelate the LSFs and extract new independent features in hash modeling.

B. Units

A circle of radius "a" rolls along the x-axis. "P" is the point on this circle of initial contact. As the circle rolls, the point "P"

traces out a curve. This is a cycloid. When the point "P" is moved to a distance "b" from the center of the rolling circle the curve traced out is a trochoid. The effects are quite different for $b < a$ and $b > a$. So, a cycloid is just a trochoid with $b = a$. For simplicity, $a = 1$ in the animated examples, and "P" starts at the origin.

The general parametric equations for a trochoid are:

$$\begin{cases} x(\theta) = a\theta - b\sin(\theta) \\ y(\theta) = a - b\cos(\theta) \end{cases} \quad (4)$$

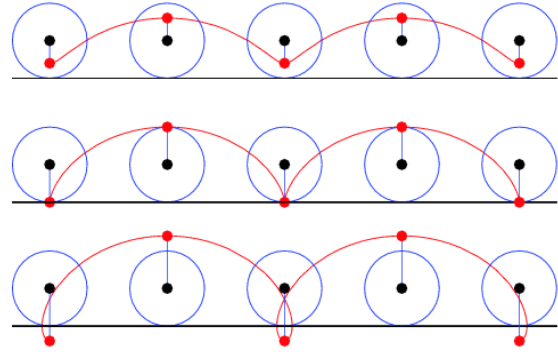


Figure 3. an trochoid graph

C. Hash Modeling

Security is essential when robust hashing is applied for content authentication. A keyed randomization scheme is introduced in hash modeling to guarantee that an unauthorized user cannot forge a valid hash without the key. The hash modeling phase contains six steps.

- 1) Seeded with a secret key, one random number sequences, I_T is generated by a uniform pseudo-random number generator (PRNG). They are used to select blocks from the SWT coefficient matrix (L).
- 2) From frame length, a pair of blocks $r_{i,1}$ and $r_{i,2}$ are chosen with the same size. $I_T(i, 1)$ and $I_T(i, 2)$ are the frame IDs of the first frame in $r_{i,1}$ and $r_{i,2}$, respectively.
- 3) $r_{i,1}$ and $r_{i,2}$ are transformed by a two-dimensional SWT components. SWT coefficients are taken with respect to robustness. The resulting features are denoted as $d_{i,1}$ and $d_{i,2}$. $d_{i,1}$ used when (i) is even and $d_{i,2}$ used when (i) is odd:
$$\begin{cases} d_{i,1} = r_{i,1}(r_{i,1} * r_{i,end}) - r_{i,2}\sin(r_{i,1} * r_{i,end}) \\ d_{i,2} = r_{i,1} - r_{i,2}\cos(r_{i,1} * r_{i,end}) \end{cases}$$
- 4) The next input rotation matrix is given below by rotating component.
$$r(j + 2, :) = d(1, :)$$
- 5) The hash components, denoted as $h_i(j)$ is decided by the sign of feature $d_{i,1}$ and $d_{i,2}$ difference of last rotation, as given in (1), where $j \in [1, F]$

$$h(i) = d_{i,1} - d_{i,2}$$
- 6) The steps from 2 to 4 are repeated for t_i times ($i \in [1, 64]$) Thus, the final hash sequence is a $F \times t$ binary matrix,

which is the content-based signature of a one second speech clip.

Besides security, random feature extraction has another advantage over the method proposed in [10]. As the pairs for comparison are randomly selected, both global and local features are used to enhance the sensitivity to tampering. The bitrate of CDWH is 0.512 Kbps, which is lower than that of rMac (5.5 KHz) [5] and of SCA (0.9 KHz) [10]. With high performance speech coders, raw speech samples could be generated with low cost. Therefore, the propose scheme of audio signal has versatility as well as low complexity.

IV. EXPERIMENTAL RESULTS

Simulation experiments have been performed in order to test, (i) the robustness of the perceptual hash for identification, where the critical behavior is the statistical spread of the audio hash function when an audio document is subjected to various signal processing attacks; (ii) the uniqueness of the perceptual hash, where the important behavior is the fact that the hashes differ significantly between two different contents. In other words, in the first case, a document is wanted to identify and its variants under signal processing attacks. In the second case, is wanted to classify documents with different contents, so that if a document should be verified, the others in the database appear as “impostors.”

A. Discrimination

BERs between 500000 pairs of hashes are calculated. The hashes are extracted from speech excerpts of different contents. The speech excerpts are randomly selected from the database described above. The normal probability plot of the measured BERs is shown in Fig. 5. If and only if the bits are independent and identically distributed (the probability of 0/1 is equal to 0.5), the bit error rate (BER) of the two hash sequences is normally distributed, with a mean of p and a standard deviation of $\sigma_0 = \sqrt{p \times (1 - p)/N} = \sqrt{1/4N}$ (N is the number of bits in a hash sequence) [8]. With the parameter in proposed scheme, the mean and standard deviation of normal distribution should be 0.5 and 0.0221, respectively. From our experimental results, we get the mean and the standard deviation, which are very close to the normal distribution values. Hence, our hash sequence is random and collision resistant. When the threshold is set to 0.1, the false positive rate is 4.2175×10^{-21} , which is much lower than the false positive rate in [10]. For each speech clip, the distance of its hash vector and the hash vector of each of the remaining 1654 speech clips was calculated by, and 498,346 distance values were obtained. A comparison between the probability density distribution of these distance values and the normal distribution shown in Fig. 4 indicates that the distance value has an approximately normal distribution. The expected value and standard deviation are $\mu = 0.4988$ and $\sigma = 0.0112$, respectively.

The FAR (calculated by (15)) varying with the threshold is shown in Table II. Fig. 4 shows a comparison of the FAR curve obtained by theoretical analysis (see Fig.4) with that obtained by the experimental values. The probability density distribution of the distance values follows the normal approximation fairly well; thus, it is verified that the threshold obtained can be used in practice with reasonable accuracy.

B. Robustness

Speech excerpts are subjected to the following content preserving manipulations.

Table II lists the average value of the BERs. Now we are to discuss the robust performance of proposed scheme.

First, the BER caused by transcoding errors is a little bit higher than by resampling. Second, the resulting BER is below the pre-specified threshold 0.1, that is, the proposed scheme is robust to noise addition.

Third, there is no difference between the LSFs of the original speech and those after volume change. It is due to the reason that volume amplifying and reducing do not change the vowels. Fourth, since MP3 compression is based on non-linear psychoacoustic model, MP3 re-encoding delivers higher BERs than speech transcoding. Fifth, the BER is still small even when the portion of the cropped speech reaches 35%. It is because that global feature is used to generate the hash in proposed scheme. Finally, when the frame desynchronization is large, time scaling usually delivers large BERs. That is why our proposed scheme is breakable when the time scaling is larger than 7%.

C. Statistic analysis of diffusion and confusion

Confusion and diffusion are two essential design criteria for encryption algorithms, including audio hash functions. Shannon initiated diffusion and confusion in order to conceal message redundancy [18,19]. Audio hash function, like encryption system, requires the plaintext to diffuse its effects into the whole Hash space. This means that the correlation between the message and the corresponding Hash value should be as small as possible. Diffusion means spreading out the influence of a single plaintext symbol over many audio symbols so as concealing the statistical structure of the audio file. Confusion means the utilizing of transformations to make difficult the dependence of audio file statistics. In the hash value in audio pairs format each audio symbol can be changed. Therefore, the perfect diffusion effect should be that any minute change in the initial condition leads to a 50% changes in average of the all symbols, change probability of each symbol. Regularly six statistics are defined as follows:

Minimum changed audio symbol number:

$$B_{min} = \min(\{B_i\}_1^N) \quad (5)$$

Maximum changed audio symbol number:

$$B_{max} = \max(\{B_i\}_1^N) \quad (6)$$

Mean changed audio symbol number:

$$\bar{B} = \frac{1}{N} \sum_1^N B_i \quad (7)$$

Mean changed probability:

$$P = \frac{\bar{B}}{L} \times 100 \quad (8)$$

Standard variance of the changed audio symbol number:

$$\Delta B = \sqrt{\frac{1}{N-1} \sum_{i=1}^N (B_i - \bar{B})^2} \quad (9)$$

Standard variance:

$$\Delta P = \sqrt{\frac{1}{N-1} \sum_{i=1}^N \left(\frac{B_i}{L} - P\right)^2} \times 100 \quad (10)$$

Where N is the total number of tests and B_i is the number of changed audio symbols in the i_{th} test. The advantage of the correlation-based method is that it requires about three times less computation as compared to the parametric estimation.

$$R(K) = \frac{\left(\frac{1}{N-K}\right) \sum_{i=0}^{N-K} s(i)s(i+K)}{\left(\frac{1}{N}\right) \sum_{i=0}^N s^2(i)} \quad (11)$$

TABLE II
STATISTICAL PERFORMANCE OF THE PROPOSED ALGORITHM

Frame	600	800	1000	1200	1400	1600
B	0.1142	0.0928	0.0889	0.0895	0.0897	0.0887
$\Delta\mathbf{B}$	0.1748	0.0961	0.1415	0.1009	0.0897	0.0903
P	50.397	49.097	50.82	49.113	51.216	51.005
$\Delta\mathbf{P}$	0.0365	0.0439	0.0389	0.0518	0.0322	0.0334
R	0.9195	0.8928	0.8995	0.9034	0.9205	0.9059

D. Security aspects of the audio hash functions

The security of the hash extraction becomes important in audio authentication schemes. One common way to provide hash security is to devise a key-based scheme such that for two different keys, K_1 and K_2 , the resulting audio hash functions become totally independent. Thus the probability of collision should be minimized, it is necessary to guarantee that two distinct inputs yield different audio hash functions and that the hash sequences are mutually independent. Notice that secure fingerprinting requires that the pirate should not be capable of extracting the hash value of the content without knowledge of some secret key. This would, for example, allow him to change the content while preserving the hash, that is, find a collision which would circumvent any hash-based authentication mechanism being used. As another example, it could also enable him to manipulate the bits while preserving the content and yet change the hash. This would be done, for example, when a pirate may want to avoid being detected by a copyright controller for unauthorized use of some content. One way to arrive at a key-based audio hash function is to project the resulting hash sequences onto key-dependent random bases. Another scheme would be to subject the analog hash sequence to random quantization [20]. In this scheme, the hash sequence is quantized using a randomized quantizer, and the quantizer itself becomes the source of randomness in the audio hash function's output. A third scheme could be based on random permutation of the observation frames with possible overlaps.

Thus a key-based sequence of visiting positions is generated and translated in saccades the frame window according to this sequence (remembrance that we used 25-millisecond windows with 50% overlap).

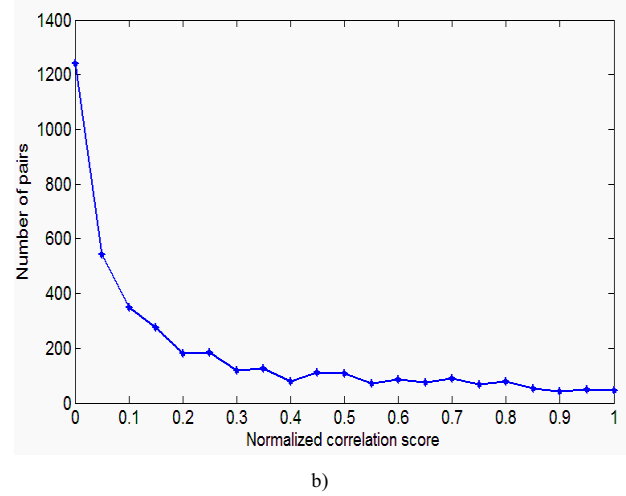
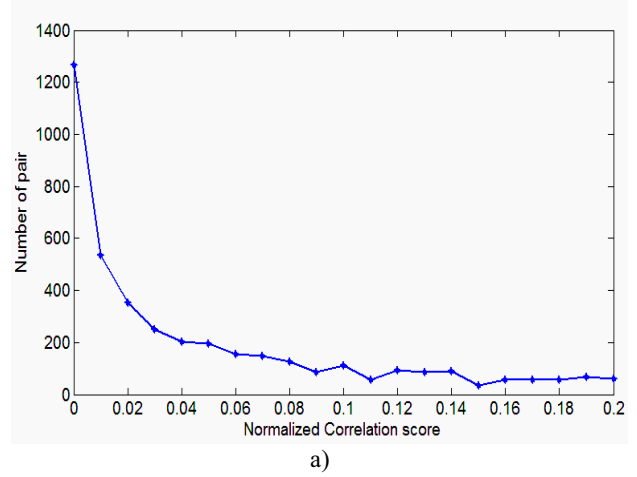


Figure 4. Histograms of the difference of the audio hash functions with 1000 speech records: (a) hashes of the different objects (solid line) and those of the attacked versions of the same object (dashed line); (b) hashes obtained from the same object with different keys.

E. Hash Matching

The problem of hash matching can be formulated as the hypothesis testing using the audio hash function $H(\cdot)$ and the distance measure $D(\cdot, \cdot)$.

L0: Two audio clips a_1, a_2 are from the same clip if

$$D(H(a_1), H(a_2)) < \tau \quad (12)$$

L1: Two audio clips a_1, a_2 are from different clip if

$$D(H(a_1), H(a_2)) \geq \tau \quad (13)$$

Where τ is a predetermined threshold, which can be obtained for a given false accept rate (FAR). FAR, denoted by R_{FA} , is

the probability that L_0 is accepted when L_1 is true. In the proposed scheme, the square of the Euclidean distance is utilized to measure the distance between any two hash vectors h_1 and h_2 .

$$x = D(h_1, h_2) = \frac{1}{L_h} \sum_{n=1}^{L_h} [h_1(n) - h_2(n)]^2 \quad (14)$$

Where L_h is the length of the hash vector. By the central limit theorem, the above distance measure has a normal distribution if L_h is sufficiently large and the contributions in the sums are sufficiently independent. Assuming that the distance measure can be approximated as the normal distribution $N(\mu, \sigma)$, the FAR is given as

$$R_{FA} = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\tau} \exp\left[-\frac{(x - \mu)^2}{2\sigma^2}\right] dx \quad (15)$$

Then, for a given R_{FA} , the threshold τ can be determined by (15), theoretically.

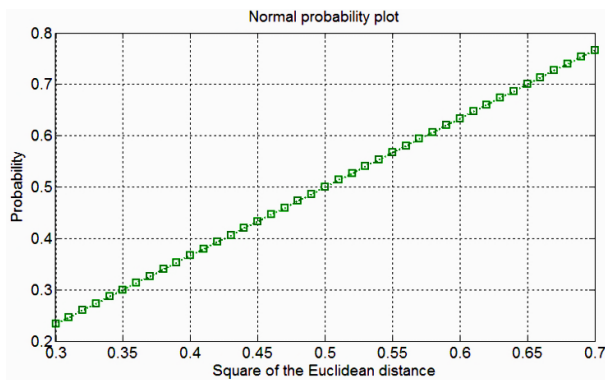


Figure 5. Comparison of probability density distribution of the distance values plotted as '□' and normal distribution.

TABLE II
FAR varying with threshold

τ	R_{FA}
0.1	4.2175×10^{-21}
0.15	2.3297×10^{-17}
0.2	4.2306×10^{-15}
0.25	7.7690×10^{-12}
0.3	3.5532×10^{-7}

V. CONCLUSION

In this paper, a novel keyed robust audio hash function is proposed. The linear spectrum frequencies are adopted for hash generation. Stationary wavelet transform is introduced to decorrelate the LSFs and enhance the discriminative capacity. As proposed robust hash function is key-dependent with high collision resistance, it could be used in multimedia authentication system. Experimental results confirmed that the proposed audio hash function is extremely robust against

speech noise addition, resampling, transcoding and other modifications. Compared with previous audio hashing, the proposed scheme is increased the tampering localization accuracy from sentence-level to alphabetic-level. For a reliable audio hash function, the feature extracted should be both discriminative and robust. In this letter, linear prediction analysis and non-negative matrix factorization were investigated for audio hash function. Linear prediction analysis was performed to extract the frequency shaping attributes of the verbal territory to realize the perceptual robustness of the proposed scheme. Experimental outcomes demonstrated the efficiency of the proposed audio hash function in terms of discernment and conceptual robustness.

Acknowledgment

This paper is supported by Iranian Research Institute for ICT (ITRC).

REFERENCES

- [1] J. S. Seo, J. Haitsma, T. Kalker, and C. D. Yoo, "A robust image fingerprinting system using the Radon transform," *Signal Processing: Image Communication*, vol. 19, no. 4, pp. 325-339, 2004.
- [2] P. Cano et al., "A Review of Audio Fingerprinting," *J. VLSI Signal Process.*, vol. 41, no. 3, 2005, pp. 271-284.
- [3] A. Ramalingam and S. Krishnan, "Gaussian Mixture Modeling of Shorttime Fourier Transform Features for Audio Fingerprinting," *IEEE Trans. Inf. Forensics Security*, vol. 1, no. 4, 2006, pp. 457-463.
- [4] M. Park, H. Kim, and S.H. Yang, "Frequency-Temporal Filtering for a Robust Audio Fingerprinting Scheme in Real-Noise Environments," *ETRI J.*, vol. 28, no. 4, 2006, pp. 509-512.
- [5] Y. Jiao et al., "Key-Dependent Compressed Domain Audio Hashing," *Proc. ISDA*, 2008.
- [6] Y. Jiao, Q. Li, and X. Niu, "Compressed Domain Perceptual Hashing for MELP Coded Speech," *Proc. IAHMSP*, 2008, pp. 410-413.
- [7] D.D. Lee and H.S. Seung, "Learning the Parts of Objects by Nonnegative Matrix Factorization," *Nature*, vol. 401, no. 6755, 1999, pp. 788-791.
- [8] T. Kalker, J. Haitsma, and J. Oostveen, "Robust audio hashing for content identification," in *Proc. International Workshop on Content Based Multimedia Indexing (CBMI '01)*, Brescia, Italy, September 2001.
- [9] M. K. Mihcak and R. Venkatesan, "A perceptual audio hashing algorithm: a tool for robust audio identification and information hiding," in *Proc. Information Hiding*, pp. 51-65, Pittsburgh, Pa, USA, April 2001.
- [10] R. Tucker, "Voice activity detection using a periodicity measure," *IEE Proceedings-I: Communications, Speech and Vision*, vol. 139, no. 4, pp. 377-380, 1992.
- [11] Mark J. Shensa, *The Discrete Wavelet Transform: Wedding the A Trous and Mallat Algorithms*, *IEEE Transaction on Signal Processing*, Vol 40, No 10, Oct. 1992.
- [12] M. Holschneider, R. Kronland-Martinet, J. Morlet and P. Tchamitchian. *A real-time algorithm for signal analysis with the help of the wavelet transform*. In *Wavelets, Time-Frequency Methods and Phase Space*, pp. 289-297. Springer-Verlag, 1989.
- [13] Mallat, S., "Zero-crossing of a wavelet transform", *IEEE Trans. on Information Theory*, 37(4): 1019-1033, (1991).
- [14] Berman, Z. and Baras, J. S., "Properties of the multiscale maxima and zero-crossings representations", *IEEE Trans. on Signal Processing*, 41(12): 3216-3231, (1993).
- [15] Kadambe, S., and Boudreaux-Bartels, G., "Application of the wavelet transform for pitch detection of speech signals", *IEEE Trans. on Information Theory*, 38(2): 917-924, (1992).

Application of geometrical algebra to neural computations.

Irakli Rodonaia, Vakhtang Rodonaia
Faculty of Computer Technologies and Engineering
International Black Sea University
Tbilisi, Georgia
irakli.rodonaia@ibs.edu.ge

Abstract- Geometrical algebra has been proved to be a powerful mathematical language for neural computations. An overview of some approaches in this area is given in the paper. Applications of the geometrical algebra methods to the geometrical transformations in computer graphics, robot vision and inverse kinematic problem are considered.

Keywords- neural computations; geometrical algebra; multivectors; geometric product; spinor neuron; spinor representation of rotation; computer graphics; robot vision; inverse kinematic problem

I. Introduction

Traditionally neural networks are built on the basis of neurons, whose inputs, outputs and activation functions have values in the area of real-valued numbers \mathbf{R} . However, this approach has some disadvantages: in many scientific and applied fields (signal processing, multidimensional time series theory, pattern recognition, robotics, system modelling, computer animation, etc.) much more sophisticated data structures (complex numbers, quaternions, multidimensional objects of complex structure) are used.

II. Statement of the problem

Let us consider, for example, the task of neuron learning (for the tasks where complex numbers are intensively used, for example, the signal processing or adaptive control) the mapping $f: \mathbf{x} \rightarrow \mathbf{y}$, where $\mathbf{x} = x_1 + ix_2$ and $\mathbf{y} = y_1 + iy_2$, $\mathbf{x}, \mathbf{y} \in \mathbf{C}$ – complex numbers [1]. This corresponds to learning a weight $w \in \mathbf{C}$ so that $w\mathbf{x} = \mathbf{y}$, that is to learn a complex multiplication. Instead of the real neuron this operation can be performed by so called complex neuron (Fig.1):

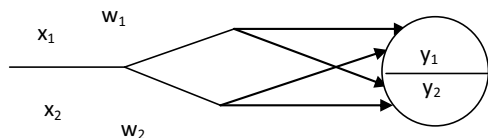


Figure 1. Complex neuron

Here the geometric neuron perceives a complex number on its input as a single whole and yields on the output also a complex number a single whole as well. The geometric neuron operates with weights w_1 and w_2 also as with a single complex number.

Now let us consider the same complex multiplication using real-valued neurons. Obviously this will require 2 real-valued neurons to compute y_1 and y_2 , respectively. Then it is necessary to define the weight $W \in \mathbf{R}(2)$, such that the condition $(x_1, x_2)W = (y_1, y_2)$, $\mathbf{x}, \mathbf{y} \in \mathbf{R}^2$ be satisfied. Here W represents a linear transformation of the vector \mathbf{x} . The real-valued neuron has to find out the constraints on W which corresponds to the matrix representation of a complex number. These constraints are obviously $w_{11} = w_{22}$ and $w_{12} = -w_{21}$. The Fig.2 illustrates this multiplication using real neurons.

As one can see, in the case of using neurons adjusted to operate with complex numbers as a single wholes, instead of the neurons operating only with real-valued numbers, just half of the parameters (one input for a complex neuron instead of two inputs for a real neuron, one output for a complex neuron instead of two outputs for a real-valued neuron, and one weight element of a complex neuron instead of two weight element for a real-valued neuron). Moreover, after having performed experiments of training a geometric and a real-valued neurons in complex multiplication it was determined that the complex neuron learns much faster: the standard square error (SSE) converges to some selected small value in much less amount of epoch than it was for real-valued neurons. Hence, as it is clear from this example, the use of complex neurons has the following advantages: the decrease of the number of parameters used (and, therefore, the decrease of the number of neurons used) and the acceleration of the training process. Naturally, it can achieved by complication of the neuron's internal structure and operations performed.

It is advisable to extend this approach to use more generalized mathematical objects, such as objects of the geometric algebra (or Clifford algebra) – multivectors (complex numbers are the particular case of multivectors).

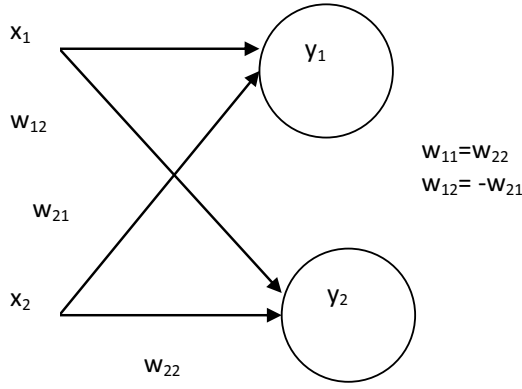


Figure 2. Multiplication using real neurons

III. Geometric Algebra Neural Networks

Informally the geometric algebra can be considered as a real vector space with the added special multiplication operation – the operation of *geometric* product. The geometric product of two vectors \mathbf{a} and \mathbf{b} is denoted as \mathbf{ab} (or $\mathbf{a} \otimes \mathbf{b}$) and can be expressed as the sum of two products $\mathbf{ab} = \mathbf{a} \cdot \mathbf{b} + \mathbf{a} \wedge \mathbf{b}$, where $\mathbf{a} \cdot \mathbf{b} \equiv (\mathbf{ab} + \mathbf{ba})/2$ is an ordinary *scalar*, or *inner*, product (real scalar), and $\mathbf{a} \wedge \mathbf{b} \equiv (\mathbf{ab} - \mathbf{ba})/2$ is an antisymmetric *outer* product of two vectors. The value $\mathbf{a} \wedge \mathbf{b}$ is neither a scalar, nor a vector. It is called a *bivector* and can be interpreted as an oriented plane segment. In the three-dimensional case a *trivector* can be served as analogue of a bivector, having a geometric interpretation of an oriented volume obtained by sweeping the area $\mathbf{a} \wedge \mathbf{b}$ along the vector \mathbf{c} . The further generalization to the spaces of higher dimension are multivectors formed by outer product of a set of independent vectors. Any multivector can be decomposed in the basis of scalars, vectors, bivectors, trivectors, etc.:

$$e_0 \equiv I, \{e_i\}, \{e_i \wedge e_j\}, \{e_i \wedge e_j \wedge e_k\}, \dots, e_1 \wedge e_2 \dots \wedge e_n \equiv I.$$

In the three-dimensional case the multivector basis of the algebra consists of 8 elements: the unit scalar, three mutually orthogonal unit vectors (oriented planes), and a trivector – an oriented unit volume associated with an imaginary unit:

$$\{e_0, e_1, e_2, e_3, e_{12}, e_{23}, e_{31}, e_{123}\},$$

where $e_{ik} = e_i \wedge e_k$ are unit bivectors, and $e_{123} = e_1 \wedge e_2 \wedge e_3 = I$ represent a unit trivector. Advantages of the geometric algebra are: it allows heterogeneous elements to be multiplied and unified, interrelations between them to be explored and complex transformation of elements to be performed using very efficient and fast operations.

Let us consider a neuron which can operate on inputs, outputs and hidden layer with multivectors. In real-valued neurons vectors are multiplied by weights using a *scalar* product. For geometric neurons scalar products are replaced with *geometric* product (Fig.3).

Here, however, correct encoding of input, weight, and output multivectors is required. For example, in three-dimensional case (which is used for three-dimensional

geometrical transformations) e_1, e_2 and e_3 are used for a basis in the space \mathbf{R}^3 .

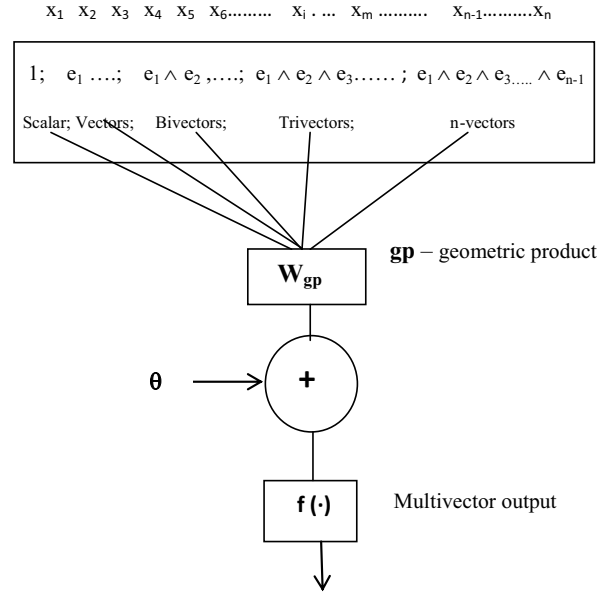


Figure 3. Geometrical neuron

The geometric algebra will have in this space the basis $I, e_1, e_2, e_3, e_1e_2, e_1e_3, e_2e_3, e_1e_2e_3$. To convert any three-dimensional vector (x_1, x_2, x_3) into the corresponding multivector \mathbf{p} it is necessary to multiply corresponding components of the vector (for example, to obtain the coefficient at the basis bivector e_1e_2 it is necessary to multiply components x_1 and x_2 , to obtain the coefficient at the basis trivector it is necessary to multiply components x_1, x_2 and x_3). The same procedure is used to encode weight multivectors and interpret output multivectors.

For geometric neuron networks the back error propagation algorithm used to train networks (in the case of supervised learning) needs to be adapted to multivector values. We have used and programmed only one of the many possible learning rules – one with component-wise sigmoid activation function to train geometric neural networks [2]. Weights w_{ij} between the i^{th} node in layer l and the j^{th} node in layer $(l+1)$ is updated by: $w_{ij} = w_{ij} + \eta \delta_j \otimes x_i^*$.

For the j^{th} node in the output layer: $\delta_j = (t_j - o_j) \odot f'(y_j)$.

For the j^{th} node on the internal (hidden) layer: $\delta_j = (\sum_k w_{jk}^* \otimes \delta_k) \odot f'(y_j)$

Here $*$ denotes geometric conjugate (by analogy with conjugate of complex numbers the meaning of conjugating multivectors is the inversion of the multivector basis with respect to the coordinate origin, as a result all multivectors change their direction for the opposite one), \otimes denotes geometric product, \odot denotes component-wise multiplication, t_j, y_j and o_j are respectively, the target and

computed network output (y_i for internal layers and o_j for the output layer) for the j^{th} output node, η is the learning rate (a real value between 0 and 1).

Learning capabilities of geometric neural networks were examined using this algorithm (corresponding real-valued neural networks were also examined for the purpose of comparison). Various linear functions were approximated by neural networks of the following configurations:

- geometrical neural network of topology 1-1-1 (i.e. 1 input geometrical neuron, 1 geometrical neuron in hidden layer and 1 output geometrical neuron) and of topology 1-2-1)
- corresponding real-valued neural network of the topologies 8-8-8 and 8-16-8.

Experiments have proved that both types of geometric networks learn much faster than real-valued ones. Besides, geometric neural networks require less weights and neurons and, as a result, less computing.

Experiments with geometric transformations (which are important in computer graphics and robot vision) were also run. The task was to learn the 3D similarity transformation (composed by a rotation with some angle, a translation with some value and a dilatation (scaling) by some factor). To perform a rotation in \mathbf{R}^3 (mapping a vector to another vector) the two-sided geometric (*spinor*) product is used (Fig.4). Here $(\mathbf{w}^*)^{-1}$ is the inverse conjugated \mathbf{w} . So, spinors are representing weights of the input of a neuron. The spinors of \mathbf{R}^3 are quaternions (quaternionic spinor) and perform a rotation in \mathbf{R}^3 . The geometric product is now the quaternion product and instead of the vector $\mathbf{x} \in \mathbf{R}^3$, we have to use its representation as a quaternion. Such a neuron (performing a spinor product) computes an orthogonal transformation by using *only* one weight \mathbf{w} . Only in the case of a spinor product we use only one neuron with one weight for computing an orthogonal transformation. Obviously, the geometrical spinor neuron has only half as many parameters (and half as many arithmetic operations) as the geometrical neuron because only the even components of the weight multivector are used.

With the following experiments the strength of the model of single quaternionic spinor neurons in comparison with multiple real neurons were tested. The transformation that should be learned was a composition of a Euclidean 3D rotation about -60° about the axis $[0.5, 0.5, 0.5]$, a translation about $[0.2, -0.2, 0.3]$. Linear transformations were performed by a real-valued single layered perceptron with 2 neurons and 4 weights and by one quaternionic spinor neuron (Fig.5- Fig.8). Experiments showed [Fig.9] that the quaternionic spinor neuron converges much faster than the real neurons than in the case of a real-valued single layered perceptron. The real neurons have to learn the matrix representation of the quaternionic multiplication. Due to that fact, it was impossible to drop the SSE < 0.00001 for the real neurons. Thus, there exists already a numerical boundary value of reachable accuracy for the computation with real neurons.

Such effective learning capabilities of geometric neurons can be explained, in our opinion, by the following factors.

Information being processed is represented by multivectors containing scalar, vector, bivector, trivector, etc. components. These components reflect intrinsic nature of information and are interrelated by some hidden, unknown before learning process links. During the learning process geometric neurons reveal these links and, based on properties of geometric algebra's operations, make appropriate changes in the values of weights, which implies acceleration of learning process. However, as we could observe from the experiments made, geometric neurons are effective only when multidimensional data being processed have a distinct geometric structure.

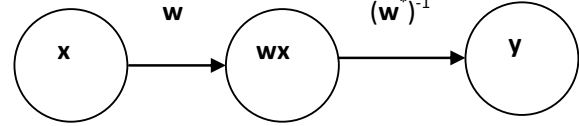


Figure 4. Spinor product.

Another approach using so called *spinor representation* of three-dimensional generalized rotations was implemented and tested [3-4]. In this approach the correspondence between set of *unitary* 2x2 matrices C in spinor space and set of real proper orthogonal rotation 3x3 matrices A in three-dimensional space is established. The method allows any vector \mathbf{x} to be rotated to the vector \mathbf{y} around some center point by using

$$\mathbf{y} = \bar{C}^T \mathbf{x} C,$$

where \bar{C}^T is the conjugated transposed C .

Vectors \mathbf{x} and \mathbf{y} have to be preliminary converted into 2x2 matrices of Hermitian functionals (consisted of so called *spinor components* [4]). The method in [4] allows Euler angles corresponding to the given rotation to be calculated. This is an important advantage because in many areas (e.g., the inverse kinematic problem for robot manipulators and computer animation) Euler angles are to be determined. However, there was no method of determining Euler angles if only coordinates of starting and final points for rotating vectors were given. Experiments to determine Euler angles for various rotation operations in inverse kinematic problem were performed. Network configuration was the same as in the experiments with geometrical neurons. Experiments showed that, despite relatively complicated organization of neurons (caused by the need to represent vectors in the form of Hermitian functionals), learning time is faster and its SSE is much less than in the case of a real-valued single layered perceptron.

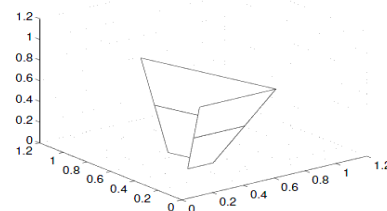


Figure 5. Training data input

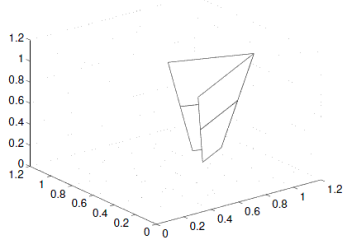


Figure 6. Training data output

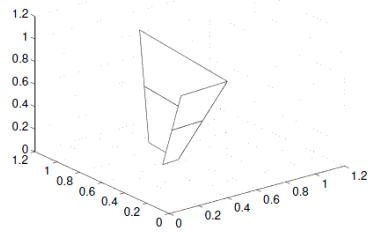


Figure 7. Testing data input

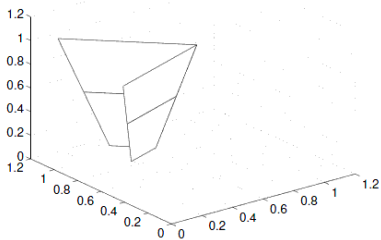


Figure 8. Testing data output

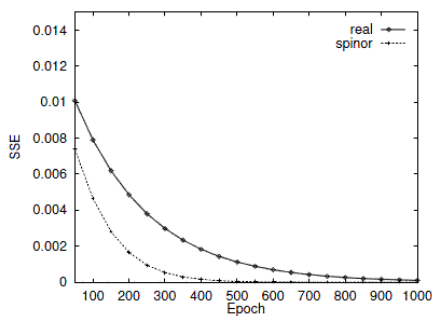


Figure 9. Convergence of the learning

- the modification of the error propagation method adopted to the multivectors
- potentials of application of geometric neural networks to linear transformation in computer graphics and robot vision are shown.
- results of experiments with linear transformation performed by real-valued and spinor neurons are shown
- possibility of using less computing resources and perform less computations
- the new approach of building neural networks with neurons using spinor representation of spatial rotations
- the possibility of solving important problem of determining Euler angles in inverse kinematic problem

REFERENCES

- [1]. E.B.Corrochano, S.Buchholz, "Geometric neural networks", Lecture Notes in Computer Science, 1997, Volume 1315, Algebraic Frames for the Perception-Action Cycle, pp. 379-394, download: <http://www.springerlink.com/content/w1w2720lu2011910/fulltext.pdf>
- [2]. Qing Yi, "Learning rules for low-dimensional Clifford neural networks", Master Thesis, Computer Science Department, Portland State University, 2004
- [3]. A.A. Milnikov, A.I. Prangishvili, I.D. Rodonaia, "Spinor model of generalized three-dimensional Rotations", Automation and Remote Control 2005, vol. 66, No6, pp. 876-882
- [4]. V.I. Rodonaia, "Development of new methods for computer modeling of spatial rotation dynamics", PhD Thesis, Tbilisi, Georgian Technical University, 2012

CONCLUSIONS

In the paper the following issues are considered:

- the advantages of using neural networks based on geometric algebra for multidimensional problems
- the fundamentals of geometric algebra.
- the structures of neural networks based on geometric and spinor neurons

The adaptive method of decision making in problems of motion terminal control

Vakhtang Rodonaia

Faculty of Computer Technologies and Engineering
International Black Sea University
Tbilisi, Georgia
vrodoniaia@ibsu.edu.ge

Abstract - The new adaptive method of decision making in problems of terminal control is proposed. Unlike the traditional program methods which are characterized by lack of feedback, the proposed method provides a continuous control over the current state of the controlled object. This requires measurements of controlled variables and corresponding corrections of them to provided the desired development of the terminal control process. The adaptive method is completely described by determining control variables and boundary conditions. Three particular cases met in practice are considered.

Keywords- decision making, terninal control, variational problem, adaptive control, controlled variables, feedback, boundary conditions

I. INTRODUCTION

In last years new problems in automatic control theory arose. Space vehicles, for instance, require minimal fuel consumption or minimal heating during the descent from the orbit and passage through the atmosphere. Such problems as robots' arms motion control, air traffic control over large airports, soft landing of space satellites, control of physical and technological processes in CAD/CAM systems, animation in computer graphics and many other problems made it extremely significant to focus attention to the methods of terminal control of objects, since these methods allow us to achieve a given phase state of the object at a given moment of time. In other words, we can, for instance, to move the object to a chosen point of the space with a given velocity vector within the desired time [1-3].

Basic idea of the terminal control is as follows. Let us consider one-dimensional motion of a controlled object, the coordinate of which is γ . It is obvious that its motion is described by the following system of differential equations

$$\begin{aligned} \dot{V} &= \frac{1}{m} \left(\sum_{i=1}^n F_i + \sum_{j=1}^k f_j \right); \\ V &= \dot{\gamma}, \end{aligned} \quad (1)$$

where V is the motion velocity of the controlled object under consideration; F_i ($i=1, 2, \dots, n$) are the projections of uncontrolled forces on the direction of motion, i.e. on the γ -axis; f_j ($j = 1, 2, \dots, k$) are the projections of controlled forces

on the direction of motion, i.e. on the γ -axis; m is the object mass. Uncontrolled forces may include, for example, all perturbations generated by the environment in which the motion takes place.

The terminal state control problem is formulated as follows. Given the initial phase state of the object $\gamma_0, \dot{\gamma}_0$, it is required to transfer it – within time T - to the terminal state $(\gamma_f; \dot{\gamma}_f)$. Uncontrolled forces are functions of time t , the coordinate γ and velocity $\dot{\gamma}$ - $F_i = F_i(t, \gamma, \dot{\gamma})$, while controlled forces, in addition to being all these functions, are also functions of the controlling parameter α ¹ - $f_j = f_j(t, \gamma, \dot{\gamma}, \alpha)$. Note that the parameter α is frequently the position of the controlling element and may be a function of time. The traditional approach to the solution of the above-stated motion control problems consists in finding the functions $f_j = f_j(t, \gamma, \dot{\gamma}, \alpha)$ for which solutions of system (1) satisfy, on the time interval $[0; T]$, the corresponding boundary conditions. The uniqueness of a solution is obtained by using an additional condition requiring that solutions must supply an extremum to some specially chosen functional. Such an additional condition is frequently the requirement for a control time minimum (quick action maximum) or an energy minimum of controlling forces.

Assume that the above said conditions are taken into account in the form:

$$\dot{\alpha}(t) = k_c (\varphi(t) - f(t, \gamma, \dot{\gamma}, \alpha(t))) \quad (2)$$

where $\varphi(t)$ is some function of controlling forces, k_c - the coefficient of proportionality.

Assume also that a relation between the controlling parameter $\alpha(t)$ and the value of the current (measured) force $f = f(t, \gamma, \dot{\gamma}, \alpha(t))$ can be written in the form of an inertia element of first order

$$\dot{f} = (k_f \alpha(t) - f). \quad (3)$$

The control process is therefore described by means of the system of differential equations (1)-(3). Knowing the

¹ For example, in the case of jet engines the throttle may play the role of a controlling parameter.

synthesized function of controlling forces $\boldsymbol{\varphi}(t)$, we can transfer object from the initial state $\boldsymbol{\gamma}(t_0); \dot{\boldsymbol{\gamma}}(t_0)$ to the terminal state $\boldsymbol{\gamma}(t_f); \dot{\boldsymbol{\gamma}}(t_f)$. However, here we encounter a difficulty caused by the necessity to measure controlling forces. This, obviously, can be done if these forces are separated from controlled forces during the object motion. Solutions obtained in this manner are of program character (the control system is open). Unfortunately, from the practical standpoint, the latter is an unsolvable problem and leads to the instability of the realized motion because of the unforeseen influence of uncontrolled forces. This circumstance requires the development of adaptive methods that demands a different approach: it is necessary to keep a continuous control over the current state of the controlled object which requires respective measurements to be taken. So, *the decision about the further development of the process must be made at every step*. Here we have to note, that in [1] it is assumed (without proving) that the control function $\boldsymbol{\gamma}(t)$ can be represented as a certain degree polynomial with unknown coefficients. Unlike, in [3-4] we suggested the rigorous formal derivation of the shape for controlled polynomial on the base of solution of a variational problem.

II. STATEMENT OF THE PROBLEM

Let us take into account the fact that any change of controlling forces brings about a change of uncontrolled forces too. All forces (uncontrolled + controlled) acting on the controlled object generate the object motion acceleration \dot{V} . It is obvious that \dot{V} can be easily measured directly and therefore we should pose the problem on the synthesis of a controlling function in the form of acceleration $\ddot{\boldsymbol{\gamma}}(t)$. Then the control process reduces to the fulfillment of the equality

$$\dot{V} = \ddot{\boldsymbol{\gamma}}(t) \quad (4)$$

where \dot{V} is the measured acceleration of the object and $\ddot{\boldsymbol{\gamma}}(t)$ is the given (synthesized) acceleration of the object.

Note that (4) is actually the equation of motion of the controlled object under the action of the controlling function $\ddot{\boldsymbol{\gamma}}(t)$ and is equivalent to (1). This is explained by the fact that the measured acceleration of the object \dot{V} takes into account changes of both uncontrolled and controlled forces

Let us assume that the relation between the given acceleration $\ddot{\boldsymbol{\gamma}}(t)$ and controlling forces $\boldsymbol{f} = \boldsymbol{f}(t, \boldsymbol{\gamma}, \dot{\boldsymbol{\gamma}}, \boldsymbol{\alpha}(t))$ is $\ddot{\boldsymbol{\gamma}} = \boldsymbol{k}\boldsymbol{f}(t, \boldsymbol{\gamma}, \dot{\boldsymbol{\gamma}}, \boldsymbol{\alpha}(t))$

where \boldsymbol{k} is the proportionality coefficient.

The synthesis of a control algorithm can be reduced to some variational problem in a phase space: given two points $(\boldsymbol{\gamma}_0; \dot{\boldsymbol{\gamma}}_0)$ and $(\boldsymbol{\gamma}_f; \dot{\boldsymbol{\gamma}}_f)$ in a two-dimensional phase space, it is required to derive the equation of a curve of this phase space that connects $(\boldsymbol{\gamma}_0; \dot{\boldsymbol{\gamma}}_0)$ and $(\boldsymbol{\gamma}_f; \dot{\boldsymbol{\gamma}}_f)$ and delivers a minimum to the next functional

$$J_F = \int_0^T \boldsymbol{f}^2(t, \boldsymbol{\gamma}, \dot{\boldsymbol{\gamma}}, \boldsymbol{\alpha}(t)) dt \quad (6)$$

The equation of the curve we want to define can be written parametrically as $\boldsymbol{\gamma} = \boldsymbol{\gamma}(t)$ and $\dot{\boldsymbol{\gamma}} = \dot{\boldsymbol{\gamma}}(t)$. Then it is obvious that to the phase curve defined in this manner there corresponds the motion trajectory from the point $\boldsymbol{\gamma}_0$ to the point $\boldsymbol{\gamma}_f$. The initial velocity at the initial moment of time $t = t_0$ is equal to $\dot{\boldsymbol{\gamma}}_0$ and at the terminal moment of time $t = T - \dot{\boldsymbol{\gamma}}_f$.

From (6) it follows that the trajectory $\boldsymbol{\gamma} = \boldsymbol{\gamma}(t)$ and $\dot{\boldsymbol{\gamma}} = \dot{\boldsymbol{\gamma}}(t)$ delivering a minimum to (6) is optimal in the sense that it minimizes energetic controlling actions.

The acceleration along the optimal trajectory is the function of phase coordinates

$$\ddot{\boldsymbol{\gamma}} = \boldsymbol{\Phi}(\boldsymbol{\gamma}, \dot{\boldsymbol{\gamma}}). \quad (7)$$

From (4) and (5) we have

$$\boldsymbol{k}\boldsymbol{f}(t, \boldsymbol{\gamma}, \dot{\boldsymbol{\gamma}}, \boldsymbol{\alpha}(t)) = \boldsymbol{\Phi}(\boldsymbol{\gamma}, \dot{\boldsymbol{\gamma}}). \quad (8)$$

Substituting (8) into (6) we obtain

$$J = \frac{1}{T} \int_0^T \boldsymbol{k}_I [\boldsymbol{\Phi}(\boldsymbol{\gamma}, \dot{\boldsymbol{\gamma}})]^2 dt = \frac{1}{T} \int_0^T [\boldsymbol{k}_I \ddot{\boldsymbol{\gamma}}]^2 dt, \quad (9)$$

where $\boldsymbol{k}_I = \frac{1}{\boldsymbol{k}}$.

Functional (9) belongs to the type of functionals that contain derivatives of second order and therefore its corresponding Euler equation can be written in the form

$$\frac{d^2 \ddot{\boldsymbol{\gamma}}}{dt^2} = 0. \quad (10)$$

Solution (10) is a third order polynomial

$$\boldsymbol{\gamma} = \boldsymbol{C}_0 + \boldsymbol{C}_1 t + \boldsymbol{C}_2 \frac{t^2}{2} + \boldsymbol{C}_3 \frac{t^3}{6} \quad (11)$$

The boundary conditions are:

$$t=0; \boldsymbol{\gamma} = \boldsymbol{\gamma}_0; \dot{\boldsymbol{\gamma}} = \dot{\boldsymbol{\gamma}}_0, \quad (12)$$

$$t=T; \boldsymbol{\gamma} = \boldsymbol{\gamma}_f; \dot{\boldsymbol{\gamma}} = \dot{\boldsymbol{\gamma}}_f \quad (13)$$

These four conditions are sufficient for defining four constants $\boldsymbol{C}_i (i=0,1,2,3)$ contained in (11), which completely defines an optimal trajectory. Below we will consider some particular cases defined by various values of the boundary conditions (12) and (13). (5)

II. ADAPTIVE METHOD OF DECISION MAKING IN TERMINAL CONTROL PROCESS.

A. Bringing Problem.

First let us consider the *bringing* problem. The problem is defined by the following boundary conditions:

$$t=0; \boldsymbol{\gamma} = \boldsymbol{\gamma}_0; \dot{\boldsymbol{\gamma}} = \dot{\boldsymbol{\gamma}}_0, \quad (14)$$

$$t=T; \gamma = \gamma_f \quad (15)$$

Conditions (14) and (15) mean that the object should be transferred from the initial state $\gamma = \gamma_0$ and $\dot{\gamma} = \dot{\gamma}_0$ to the state $\dot{\gamma} = \dot{\gamma}_f$ and at that its motion velocity should be arbitrary. For problems of this kind, the given boundary conditions (14) and (15) are supplemented by the so-called natural boundary condition which in our case looks like [5]:

$$G\ddot{\gamma} - \frac{d}{dt}G\dot{\gamma} = 0 \quad (16)$$

where $G = k\ddot{\gamma}^2$.

After some elemental transformation [4] we obtain expressions for C_i ($i = 0, 1, 2, 3$):

$$C_3 = 0; C_2 = \frac{2(\gamma_f - \gamma_0)}{T^2} - \frac{2\dot{\gamma}_0}{T}; C_1 = \dot{\gamma}_0; C_0 = \gamma_0 \quad (17)$$

Substituting (17) into the first and the second derivative of (11), we obtain the following expressions for an optimal trajectory in the phase space:

$$\gamma = \left(\frac{2(\gamma_f - \gamma_0)}{T^2} - \frac{2\dot{\gamma}_0}{T} \right) \frac{t^2}{2} + \dot{\gamma}_0 t + \gamma_0 \quad (18)$$

$$\dot{\gamma} = \left(\frac{2(\gamma_f - \gamma_0)}{T^2} - \frac{2\dot{\gamma}_0}{T} \right) t + \dot{\gamma}_0 \quad (19)$$

The acceleration (the second derivative of (18)) has the form:

$$\ddot{\gamma} = \frac{2(\gamma_f - \gamma_0)}{T^2} - \frac{2\dot{\gamma}_0}{T} \quad (20)$$

This is the law of control for the bringing problem. It means that if the acceleration of the controlled object on the time interval $[0; T]$ is assumed to be constant and equal to (20), then at the moment of time $t = T$ its state will satisfy the boundary conditions (13). However, this is an open (program) law of control, i.e. the control law without feedback. Due to the possibility of direct measurements of the acceleration of a controlled object, (19) can be transformed to the control law with feedback [1]. For this it is enough to assume the initial phase state to be the current one, i.e. to assume $\gamma = \gamma_0$ and $\dot{\gamma} = \dot{\gamma}_0$. In that case, the task fulfillment time should be assumed to be equal to the remaining time $T-t$. Then (20) takes the form:

$$\ddot{\gamma} = \frac{2(\gamma_f - \gamma)}{(T-t)^2} - \frac{2\dot{\gamma}}{(T-t)} \quad (21)$$

From (21) we see that in this case the acceleration that affects the controlled object ceases to be constant and becomes dependent on the current velocity and coordinate values of the controlled object, i.e. we have the realization of control with feedback. (Fig.1) represents the flowchart that implements such a control with feedback. Measured coordinates $(\gamma, \dot{\gamma})$ of the current state enter the block of automatic control system, where the required value of affecting acceleration is being generated. Thus, here decision on required actions at each step

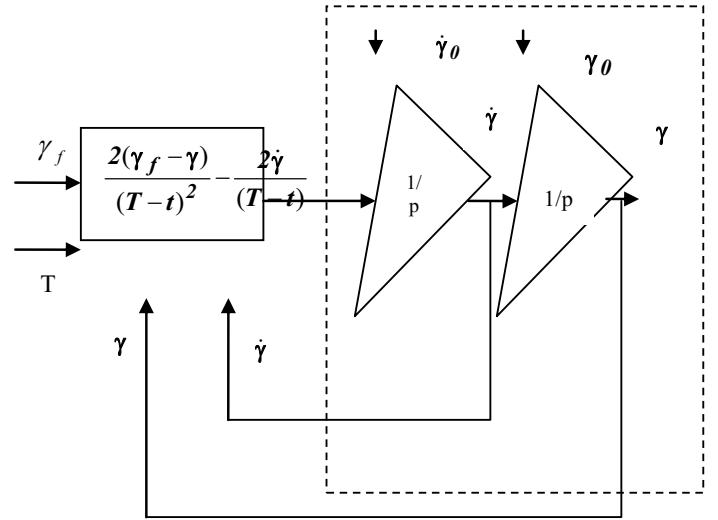


Figure 1. Control flowchart

of the control process is made. That's why the method being described is called "adaptive method of decision making".

B. Acceleration Problem

In the acceleration problem the boundary condition (13) is replaced by

$$t=T; \dot{\gamma} = \dot{\gamma}_f, \quad (22)$$

which means that in this case it is required that at the given moment of time $t = T$ the velocity of the controlled object reach the given value $\dot{\gamma} = \dot{\gamma}_f$. The coordinate may have an arbitrary value. The natural boundary condition (16) remains as before. Analogously to the bringing problem, we obtain the following values of the coefficients C_i ($i = 0, 1, 2, 3$) in the expression for the controlling program (11):

$$C_3 = 0; C_2 = \frac{\dot{\gamma}_f - \dot{\gamma}_0}{T}; C_1 = \dot{\gamma}_0; C_0 = \gamma_0 \quad (23)$$

Hence:

$$\gamma = \left(\frac{\dot{\gamma}_f - \dot{\gamma}_0}{2T} \right) t^2 + \dot{\gamma}_0 t + \gamma_0, \quad (24)$$

$$\dot{\gamma} = \left(\frac{\dot{\gamma}_f - \dot{\gamma}_0}{T} \right) t + \dot{\gamma}_0. \quad (25)$$

$$\ddot{\gamma} = \frac{\dot{\gamma}_f - \dot{\gamma}_0}{T}. \quad (26)$$

The last expression is the law of acceleration process control. It means that if on the time interval $[0; T]$ the controlled object is subjected to control (26), then at the moment of time $t = T$ its velocity will satisfy the boundary condition (22), i.e. the acceleration problem will be thereby proved.

However, this is again the program law of control and to make it self-correcting (adaptive) we proceed as in the case of the bringing problem, i.e. we replace the initial velocity

and coordinate values by the respective current values, and the moment of time T by the difference T - t:

$$\ddot{\gamma} = \frac{\dot{\gamma}_f - \dot{\gamma}}{T - t} \quad (27)$$

C. Approach Problem

The approach problem employs four boundary conditions (12) and (13) which allow us to calculate immediately the coefficients C_i ($i=0, 1, 2, 3$) in the controlling function (11):

$$\begin{aligned} C_0 &= \gamma_0, C_1 = \dot{\gamma}_0 \\ C_2 &= \frac{6}{T^2}(\gamma_f - \gamma_0) - \frac{2}{T}(2\dot{\gamma}_f + \dot{\gamma}_0) \\ C_3 &= \frac{12}{T^3}(\gamma_0 - \gamma_f) - \frac{6}{T^2}(\dot{\gamma}_f + \dot{\gamma}_0) \end{aligned}$$

But frequently it is not enough to have four boundary conditions (12) and (13) to solve the applied problems of terminal control. For example, in the case of deceleration it is not enough to assume that the terminal velocity is equal to zero: for a complete stop it is necessary that the terminal acceleration be equal to zero too. Thus, an additional boundary condition (the fifth one) related to acceleration arises:

$$\begin{aligned} t=0; \gamma &= \gamma_0; \dot{\gamma} = \dot{\gamma}_0, \\ t=T; \gamma &= \gamma_f; \dot{\gamma} = \dot{\gamma}_f; \ddot{\gamma} = \ddot{\gamma}_f. \end{aligned} \quad (29)$$

It is clear that in this case the controlling function should be taken in the form of a polynomial of fourth order containing five coefficients, of which only three are to be defined, since it is obvious that the first two coefficients satisfy the first two (initial) conditions (29):

$$\gamma(t) = \gamma_0 + \dot{\gamma}_0 t + C_2 t^2 + C_3 t^3 + C_4 t^4$$

Calculating the first and second derivatives, substituting them into the last three equations (29) and passing to the control with feedback, we obtain the relevant values of the coefficients C_i ($i=2, 3, 4$), $\gamma(t)$, $\dot{\gamma}(t)$, $\ddot{\gamma}(t)$.

The adaptive method described above was applied to the problem of spatial rotation of robot manipulator.

A. Conclusions

- the adaptive method of decision making for terminal control of motion is considered in the paper.
- terminal control methods allow us to achieve a given phase state of the object at a given moment of time
- traditional methods of terminal control have program character (the control system does not have feedback) that leads to the instability of the realized motion due to the unforeseen influence of uncontrolled forces
- adaptive methods (with feedback) must keep a continuous control over the current state of the controlled object which requires respective measurements to be taken
- decision about the further development of the terminal control process (based on measurements taken) must be made at every step.

- in the proposed adaptive methods measured coordinates of the current state enter the block of automatic control system, where the required value of affecting acceleration is being generated. Thus, here decision on required actions at each step of the control process is made.
- the synthesis of a control algorithm can be reduced to variational problem in a phase space
- the adaptive method is completely described by determining control variables and boundary conditions
- three particular cases defined by various values of boundary conditions are described

References (28)

- [1]. Batenko A.P. Control of Terminal States of Moving Objects. Moscow, Sovetskoe Radio, 1977, 225 с. (Батенко А.П. Управление конечным состоянием движущихся объектов. М.: Советское Радио, 1977, 256 с.)
- [2] Milnikov A.A., Analysis and Synthesis of Reduction Problem of Terminal Control // Problems of Mechanics, Tbilisi, 2008 №1(30), p.67-70
- [3] Milnikov A.A., New methods in Theory of Spatial Rotations Terminal Control//Proceedings of International Conference "Non-Classical Problems of Mechanics", Kutaisi, Georgia, 2007, Vol I, p.17-21
- [4] V.I. Rodonaia, "Development of new methods for computer modeling of spatial rotation dynamics", PhD Thesis, Tbilisi, Georgian Technical University, 2012
- [5] A.E.Bryson, Ho Yu-Chi, "Applied Optimal Control: Optimization, Estimation, and Control", Blaisdel Publishing Company, London, 1969

BSS parameters and their influences in GSM mobile networks

Mohammad Reza Salehifar
 Islamic Azad University
 Science and Research Branch
 Tehran – IRAN
Mr_salehifar@yahoo.com

Saeed Soleimany
 Islamic Azad University
 Qazvin Branch
 Qazvin – IRAN
Saeed.soleimany@gmail.com

Hassan Karbalaee
 Shahed University
 Tehran – IRAN
H79_karbalaee@yahoo.com

Abstract: In this paper, by review of optimizing parameters and its application in mobile networks are trying to mention effective factors, so as to improve network performance and increase the KPI (Key Performance Indicator) coefficient. In this case, software such as TEMS and Mapinfo will greatly aid and facilitate.

Keywords: GSM, BTS, BSC, TRAU, MSC

I. INTRODUCTION

A) *GSM network architecture:* Global mobile network structure is shown in figure 1(a) [1] and flowchart 1(b).

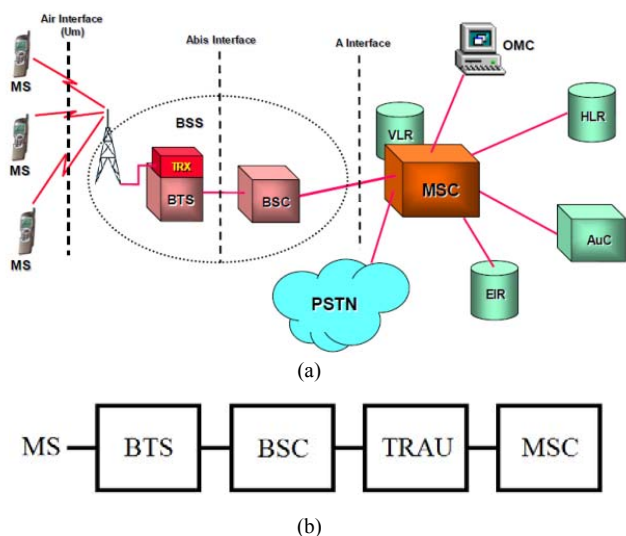


Fig 1. Global System Mobile (GSM) network structure

In this structure, Mobile Subscriber (MS) that are included from phone and SIM card interfaces through UM with the Base Transceiver System (BTS) unit that is responsible for call processing functions. Note that, in this network architecture, only link between MS and BTS is from the air and the rest of the way through a physical cable or transmission system will be closed.

B) *BTS unit:* BTS unit is tasked to process and manage calls and it is associated through the Abis interface with the central Base Station Controller (BSC). This communication

is possible through the LAPD protocol. In this network, the BTS and BSC units set so-called Base Station Subsystem (BSS). From other view, mobile networks can be divided into two main parts: BSS and NSS (Network Switching Subsystem). NSS unit encompasses all equipment related to the switching network, which can be communicated with many different parts such as HLR (Home Location Register), VLR (Visitor Location Register), AUC (Authentication Center) and EIR (Equipment Identity Register).

The BTS unit has three main sections that include: (a) combiner circuits, (b) processing units (TRX), (c) control units. The structure of this unit is identical in all Vendors and Manufacturers. Circuit combiner composed of a series of amplifier (LNA) and duplexer circuits, which are responsible for the separation and distribution of TX and RX signals. The TRX which is responsible for call processing functions includes eight time slots that the combination of channels and Voice Signaling can be defined in them. For example, for a site with a 4-4-4 combination, which is comprised of three sectors, in total, 12 frequencies can be implemented and can be covered parts of a region.

C) *BSC unit:* in fact, the BSC unit is the brain of system in mobile networks and its duty is controlling and managing of all regional BTS. The capacity of a BSC can be expressed with the number of TRX that it covers. The BSC through Abis interface can be able to control and management a BTS, from least to most alarms.

D) *TRAU (Transcode and adaptive unit) unit:* This unit, that is responsible for data compression and reopening, is a part of a BSC unit and is mounted in its rack, which will reduce the hardware size and facilitate the communication between two networks.

E) *MSC unit:* MSC has responsibility to create and switch between calls, which can include various components such as VLR that registers temporary information about subscribers. The MSC via interface A is associated with TRAU. This communication is possible through the A channels with 30 channels. In fact, a TRAU will lead a 120-channel link as input to MSC.

II. Optimization process

A) *The definition of optimization:* In general, the goal of optimization is to ensure network performance in QOS (Quality of Service) standard range. The main reasons for having done the optimization process are: a) Maintain and enhance its service quality, b) Optimum use of available resources, c) Solving network problems.

The main objective of optimizing a network is to increase the overall quality of the current network, which is defined by the coefficient of performance KPI (Key Performance Indicator). In general, the main stages of optimization simulations are available through the flowchart is shown in Figure 2.

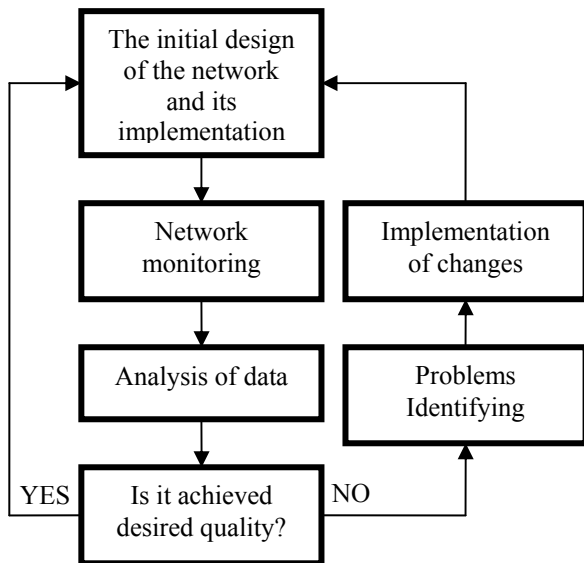


Fig 2. The main stages of optimization simulations

B) *Mobile network design:* In mobile network design, these three parameters are very high importance: 1) coverage, 2) capacity, 3) quality of service. When a mobile network is designed, first step is providing a convenient cover for the area. To achieve this goal is very important geographic location selection of the site, which can also provide coverage and not cause interference at other sites due to Overshooting. The second is to determine the traffic load, which we can assess the security channel for that area and finally, the customer will have high quality service.

In mobile networks, evaluation criteria (KPI) are considered separately for each of the BSS and the NSS. In total, the region output estimation is based on it.

III. BSS optimizing main parameters

As previously discussed, the criterion for performance assessment of a region is the coefficient of the KPI, which has separately calculation ability for each BSS and NSS units. In fact, changing this coefficient is done through parameter optimization. These parameters can be extracted in a network either through the BSC and software such as TEMS. In other words, the network quality can be determined from this method. The most important KPI indicators in the BSS including:

1) CSSR (Call Setup Success Rate)

- 2) DCR (Dropped Call Rate)
- 3) TCH Congestion
- 4) SDCCCH Congestion
- 5) HOFR (Handover Fialure Rate)
- 6) HOSR (Handover Success Rate)

Before the expression of these indices, must point to an important parameter, which is visible in the software TEMS and is effective in improving the KPI.

A) *C/I (Carrier to Interference) parameter:* This parameter indicates the amount of frequency interference between adjacent cells. Because, in mobile networks, the number of available carriers can not be held accountable by the number of setup sites in two working bands (900MHz and 1800MHz), Even with using the Reuse technique, frequency interference can not be removed. Thus, through a series of techniques must provide advantage of these limited resources for countless sites. An important method used in this section is Frequency Hopping technique That is one of the factors to increase the C/I ratio in the mobile networks. The standard C/I, in mobile networks, can be defined as $C/I > 9\text{dB}$. An example of this parameter graph can be seen in Figure 3.

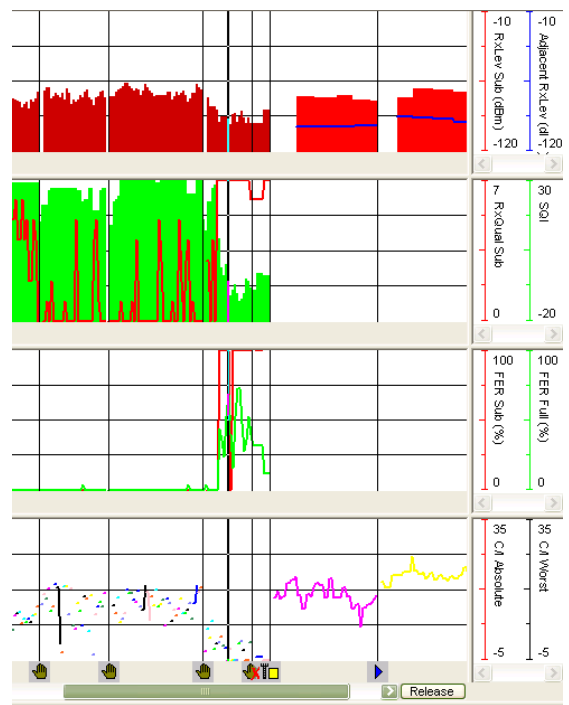


Fig 3. (a) Received signal power level (RX level), (b) quality of RX Qual, (c) the percentage error in the data frames FER, (d) C/I ratio

The frequency hopping technique in mobile network usually is applicable in two ways:

1. BBHOPP (Base Band Hopping) method
2. SFH (Synthesizer Hopping) method

Thus, a customer in network is never on a fixed frequency and frequency is spinning. For example, in first approach, for a sector with a combined 4 TRX, All conversations on the defined frequencies for each TRX, except BCCH (Broadcast Channel), was spinning and not fixed. In the

second method, one can speak through the MA list definition on subscribers' network in a frequency range on different frequencies and in circulating. In this method, the rotation algorithm and also the frequency hopping is adjustable through two parameters MAIO (Mobile Allocation Index Offset) and HSN (Hopping Sequence Number). Most important factor that increases the C/I ratio is a proper design of frequency between sectors and the frequency hopping approach can increase this parameter to an acceptable level. Two types of interference can be defined:

- 1) Co-Channel type Interaction, which takes place between two or more cells with same frequencies.
- 2) Interference with neighboring channels, which exist in one or two frequency difference units. For example, the ratio of carrier power, with frequency number 100 to carrier with frequency number 101, which is defined in the network with the C/A.

B) *CSSR parameter*: This parameter is in fact represents the number of successful calls in mobile networks. The standard rate of this parameter usually is $CSSR > 98\%$ for networks. Examples of successful call setup as shown in Figure 4.

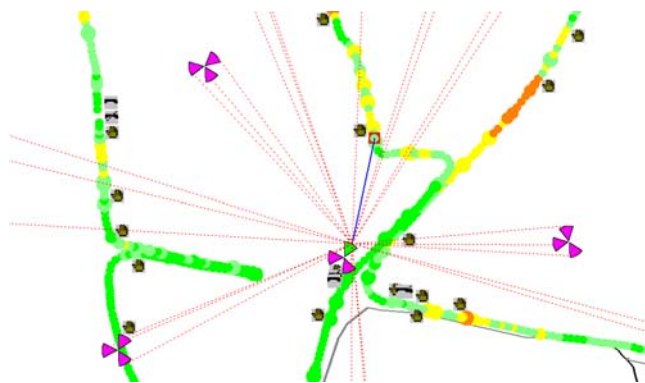


Fig 4. An example of a done Drive Test, which indicate many call setup and handovers

Reasons for the decline in this parameter can be expressed as follows:

- 1) The most important factors that can be effective in lowering the CSSR in the network, is hardware problems related to the BTS. For example, failures in one TRX unit, which can cause such a phenomenon.
- 2) The second factor could be due to lack of access network signaling channel (SDCCH Congestion). The most important signaling channel, in the mobile network is SDCCH channel, which acts on the Uplink and Downlink bidirectional.
- 3) Lack of access to free traffic channel (TCH Congestion) similarly resulted in a failure to communicate.
- 4) The next factor could be due to the VSWR (Voltage Standing Wave Ratio) phenomenon, which include loss of signal in the antenna to the BTS. Lack of proper impedance matching in the route, due to the penetration of water feeder or

antenna failure and etc., is causing this problem. These failures can be found by Site Master Device and the problem can solve.

- 5) The fifth factor can be considered as a feeder swap or sector transportation, which can be identified and solved by TEMS software through Drive test.
- 6) Lack of appropriate frequency allocation for a site is also a factor in reducing this parameter. Therefore, with proper frequency design, so that the resulting C/I were good and cluster implementation can be strengthened with this parameter.

The formula for calculating this parameter is:

$$CSSR = \frac{\text{Number of Established Call}}{\text{Number of Call Attempts}} \times 100$$

C) *DCR parameter*: This parameter is a sign of all conversations, which leads to failure in the mobile network. The standard rate in the network usually can be defined as $DCR < 2\%$. Several factors can cause this phenomenon, which include:

- 1) Hardware problems
- 2) Not defined neighborhood between two cells, which must be full duplex in the network. For cells, that a neighborhood is not defined, usually will not perform the Handover operation and finally, conversation be disconnected.
- 3) Another important factor, which can be effective in the DCR, is miss channel allocation to the moving MS from one cell to another, due to not allocate traffic channel (TCH).
- 4) Lower level of the received signal, is also important to disconnect a conversation, which may be achieved due to poor design or getting into an environment with excessive weakening.
- 5) Existing a lot of interference on the service sector and of course, the service from outside of cell area due to overshooting, will also cause failure.

Of course, all of them are visible in the TEMS software. Example of Drop Call, due to excessive interference on the sector is shown in Figure 5.

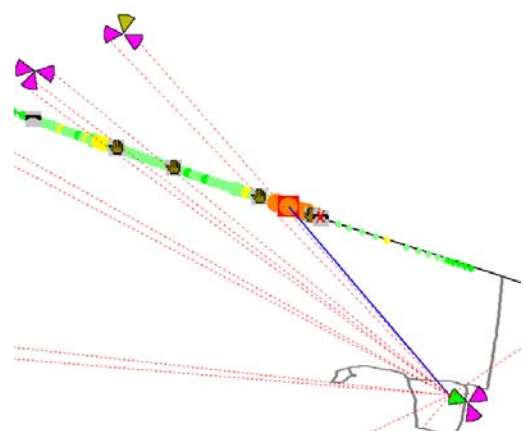


Fig 5. An example of a done Drive Test, that indicates the drop call due to interference effect

DCR parameter in the network can be calculated from following relationship:

$$CDR = \frac{\text{Number of Dropped Call}}{\text{Number of Established Call}} \times 100$$

D) *SDCCH & TCH Congestion*: Increase the value of these parameters has a direct impact on other KPI parameters, which reduces the CSSR and HOSR and this will increase the DCR and HOFR. This problem is caused by lack of cell capacity and can be solved with increased traffic or signaling channel. To increase the traffic channel, it can be first step, channel changes from Full Rate to Half Rate. However, call quality is reduced. If the continuing problem, the physical increase TRX of desired sector solves the problem. To increase the SDCCH channel, we only need to change the nature of the time slots and we define the signaling channel type.

E) *Handover Failure Rate*: As we know, in the network, Handover is the changing profile of Mobile Station (MS) or Cell ID from one area to another area. Handover is an important function in mobile communication systems. As a means of radio link control, handover enables users to communicate continuously when they traverse different cells. The major purpose of handover is to guarantee call continuity, improve speech quality, reduce cross interference in the network, and thus provide better services for mobile station (MS) subscribers. What can cause this is not done, can be stated as follows:

- 1) Insufficient capacity in neighbouring cells to accept call
- 2) low signal reception from neighbouring cells
- 3) Lack of good neighborliness definition between the cells
- 4) incorrect threshold parameters settings
- 5) a lot of interference on the target sector
- 6) Poorly designed cell plan may result in neighbouring cells being received at similar levels. This may cause confusion and therefore make cell selection more difficult, possibly resulting in handover failures.

A sample of the HOF, in Figure 6 is shown. Handover can be between two cells with the same BSC, or it will be implemented between two cells with different BSC. HOFR parameter is implemented through the following relationship. The standard rate of this parameter for the network, usually is $HOFR < 2\%$.

$$HOFR = \frac{\text{Number of Handover Failure}}{\text{Number of Handover Attempts}} \times 100$$

HOSR Parameter is also complement of HOFR and is equal to:

$$HOSR = 100 - HOFR$$

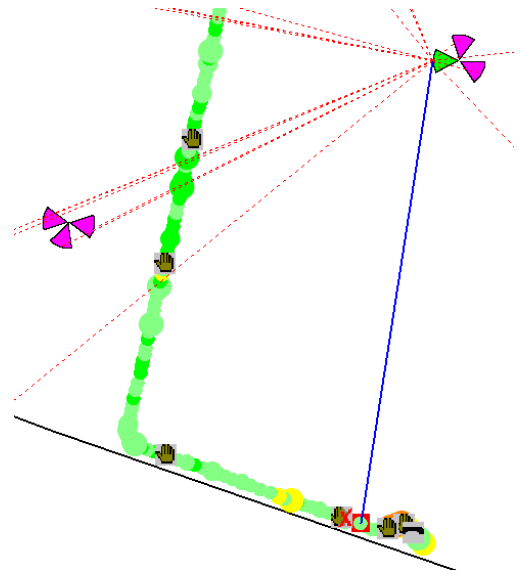


Fig 6. One sample of done Drive Test, which is indicating Handover Failure

IV. CONCLUSION

Due to the high influence of hardware problems, transmission and frequency interference than other components, it is necessary to focus more on these issues, to improve KPI Index. In fact, we solve these problems and so all indicators are improving and efficiency will be higher. With the above problems solving, focusing on each index can be paid separately to improve it.

V. REFERENCES

- [1] GSM System Overview, AIRCOM Company, 2004
- [2] GSM Network Performance Management & System Optimization, AIRCOM Company, 2004
- [3] GSM Radio Optimization Solution, Motorola, 2003
- [4] Tems Product ver9.1, Ericsson, 2004

Feature Based Iris Recognition System Functioning on Extraction of 2D Features

Arjun Agrawal
Systems Engineer
Infosys Limited
Bangalore, India
arjun.4060@gmail.com

Gundeep Singh Bindra
Department of Computer Science
SRM University
New Delhi, India
mailbox@gundeepbindra.com

Priyanka Sharma
Department of Computer Science
Global Institute of Technology
Jaipur, India
Priyanka.sharmagit@gmail.com

Abstract—In this paper a novel and simple iris feature extraction technique is proposed for iris recognition of high performance. We use one dimensional circular ring to represent iris features. The reduced and significant features afterward are extracted by Sobel Operator and 1-D wavelet transform. To improve the accuracy, this paper combines uses Euclidean Distance for classification.

Keywords:- *iris recognition; wavelet transform; probabilistic neural network; particle swarm optimization*

1. INTRODUCTION

Among all the biometric indicators, iris has one of highest levels of reliability. The human iris, an annular part between the pupil and the white sclera as shown Fig. 1(a), has an extraordinary structure and provides many interlacing minute characteristics such as freckles, coronas, stripes, etc. These visible characteristics, which are generally called the texture of the iris, are unique to individuals. Even identical twins having similar DNA, are believed to have different iris.

Iris recognition [1-13] is the process of automatically differentiating people on the basis of individuality information from their iris images. The technique can be used to verify the identity of a person when accessing a system. Due to its reliability and high precision, it is beneficial for biometric authentication system.

A typical iris recognition system can be composed into three modules: an iris detector for detection and location of iris image, a feature extractor and a matcher module.

In this paper, we focus our investigation on a new iris detector; feature extraction and representation approach to further implement an iris recognition system with low complexity and high performance.

Firstly, in order to reduce system complexity, we

use 2-D wavelet transform [14] to obtain a low resolution image and localize pupil position. By the center of pupil and the radius of pupil, we can acquire the iris circular rings. The more iris circular rings are acquired, the more information is abundant. Secondly, we segment the iris image into three parts and two parts. In each segmented iris image, iris texture is extracted as feature vector by Sobel operator. The 1-D discrete wavelet transform is adopted to reduce the dimensionality of the feature vector. In our experiments, the wavelet permits to further reduce the system complexity and obtain a discriminated feature vector. We use Euclidean Distance approach for classification problems. Finally, the combination of the novel feature extraction method and Euclidean distance classifier is evaluated on the IITK iris database for iris recognition. The experiment results present that the proposed method is well suitable for a low complex computation and low power devices.

2. DETECTION OF IRIS REGION

The iris image, as shown in Fig. 1 (a), contains not only abundant texture information, but also some useless parts, such as eyelid, pupil, etc. The iris is between the pupil (inner boundary) and the sclera (outer boundary). In order to locate the pupil, a simple and efficient method is proposed. The procedure is as following:

1. Take a raw image and apply 2-D wavelet filtering. The size of the resulting image is only quarter of the original image.
2. Compute the histogram to find the maximum peak.

The Fig. 1(b) shows that the maximum peak of histogram is the gray values of pupil region, because the pupil region is concentrated on the lower gray values. The maximum peak is set to P and the threshold T is obtained by P×W. The W is weight and the value of it is set to 1.1 in the paper. The binary image B is obtained by the original image A.

$$B(i, j) = \begin{cases} 1, & \text{if } A(i, j) > T \\ 0, & \text{if } A(i, j) < T \end{cases}$$

Because the binary image B has still some black points outside the pupil region, an estimated point is computed by a function E(i, j).

$$E(i, j) = \prod_{ii=-1}^1 \prod_{jj=-1}^1 B(i+ii, j+jj)$$

And

$$B'(i, j) = \begin{cases} B(i, j), & \text{if } E(i, j) > 4 \\ 1, & \text{otherwise} \end{cases}$$

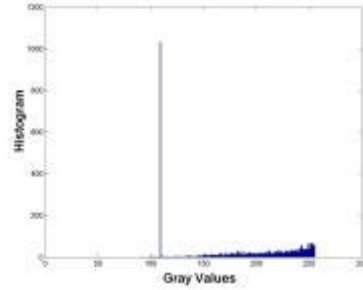
If the value of E(i, j) is greater or equal to 4, there are at least four dark points surrounding the B(i, j). Because the estimated point is located in pupil region, the estimated point is retained. Otherwise the estimated point is removed. Finally, we use the vertical projection and horizontal projection to obtain the center coordinates and the radius of pupil. The extracted pupil region is shown in Fig. 1(c).

3. Because the center coordinates and the radius of the pupil are multiplied by two, the center coordinates and the radius of the pupil are obtained in original eye image.
4. The iris circular ring (such as seen in Fig. 2 (a)) is obtained by giving a radius from the center of the pupil.

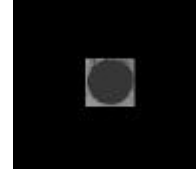
In the above mentioned procedure, the first step reduces the dimensionality of image to improve the efficiency of iris image extraction. The second and third steps provide an approach to locate the position of the pupil.



(a)



(b)



(c)

Fig 1.(a) Original image (b) The histogram of the gray values (c) Segmented pupil

We give different radius to get iris circular rings of different size. The more iris circular rings are extracted, the more information is used as features. The recognition performance is much better, but the efficiency is slightly affected. The proposed method is different from the traditional methods. The traditional methods extract a complete iris image, but the proposed method only extracts parts of the iris image for recognition. This will result in lower computational demands. In the next section, the detailed description of the iris feature extraction method will be presented.

3. IRIS TEXTURE FEATURE EXTRACTION

We extract consecutive circular rings using step 4 of iris location procedure. The more iris circular rings are extracted, the more information is used as features. The recognition performance is much better, but the efficiency is slightly affected. The proposed method is different from the traditional methods. The traditional methods extract a complete iris image, but the proposed method only extract parts of the iris image for recognition. This will result in lower computational demands.

A) Circular-derived iris blocking image

These circular rings then are stretched horizontally and accumulated, and construct a rectangular-type iris block image, shown as in Fig. 2 (b). Iris texture has abundant texture information for iris recognition. Here we elaborate a very simple and fast algorithm to extract iris feature for iris recognition. We previously proposed a novel iris feature extraction [13], but the recognition performance is not good.

In order to improve the recognition performance, the iris image is divided into three parts (see Fig. 2 (c)) and two parts (see Fig. 2 (d)). In order to compensate for a variety of lighting conditions, the segmented iris image is normalized (see Fig. 2 (e)).

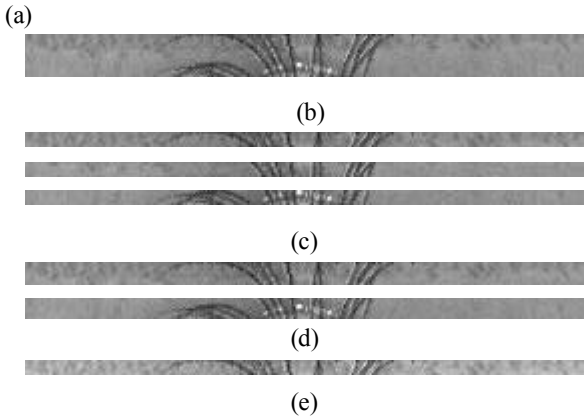
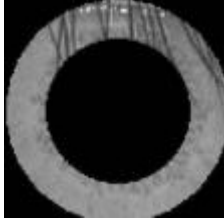


Fig.2 (a) original iris image (b) stretched iris block image;(c) iris image divided into three parts; (d) iris image divided into two parts; (e) normalized iris image

B) Sobel operator

The iris image is captured in different size from different people. It is not convenient for iris recognition, and the recognition performance is also affected. In the cause of the convenience of computation and achieving the high recognition performance, the number of captured iris circular ring from different iris image is the same. In order to enhance the texture of iris, the iris image is normalized. We adopt the Sobel operator to analyze texture shown as in Fig. 3 and the Sobel mask S_x is as following:

$$S_x = \begin{bmatrix} -1 & 0 & 1 \\ -2 & 0 & 2 \\ -1 & 0 & 1 \end{bmatrix}$$



Fig 3. Iris image after Sobel operator

C) Vertical projection

We adopt vertical projection to obtain 1-D energy profile signal and to reduce system complexity. In order to concentrate the energy, every row is accumulated as energy signal.

Let G be a segmented iris image of size $m \times n$, m is the number of iris circular ring, and n is pixels of each iris circular ring.

$$G = \begin{bmatrix} g_{1xn} & \dots & g_{1xn} \\ \vdots & \ddots & \vdots \\ g_{1xn} & \dots & g_{1xn} \end{bmatrix}$$

After vertical projection, the 1-D energy signal S is obtained.

$$S = [s_1 \ \dots \ s_n]$$

The m is much smaller than the n . Thus, the information of iris texture after vertical projection is more than the information after horizontal projection. So, we adopt the vertical projection to extract the 1-D energy signal.

D) Wavelet transform

The wavelets [14] to signal and image processing have provided a very flexible tool for engineers to apply in various fields such as speech and image processing. In an iris recognition system, the 2-D wavelet transform is only used for preprocessing. The preprocessing helps to reduce the dimensionality of feature vector and to remove noise. Nevertheless, the computational complexity is comparatively high. Thus, the paper proposes 1-D wavelet transform as filter to reduce the dimensionality of feature vector, and it can further reduce the computational complexity.

The wavelet is constructed from two-channels filter bank as shown in Fig. 4 (a). In wavelet decomposition of 1-D signal, a signal is put through both a low-pass filter L and a high-pass filter H and the results are both low frequency components $A[n]$ and high frequency components $D[n]$. The signal $y[n]$ is reconstructed by the construction filters H and L .

The wavelet filters are used to decompose signal s into high and low frequency by convolution.

$$D[n] = \sum_{k=-\infty}^{\infty} s[k] \cdot H[n-k] \quad D = s, H$$

$$A[n] = \sum_{k=-\infty}^{\infty} s[k] \cdot L[n-k] \quad A = s, L$$

In order to construct multi-channel filter, we can cascade channel filter banks. Fig. 4 (b) represents a 3-level symmetric octave structure filter bank. This is an important concept from multi-resolution analysis (MRA).

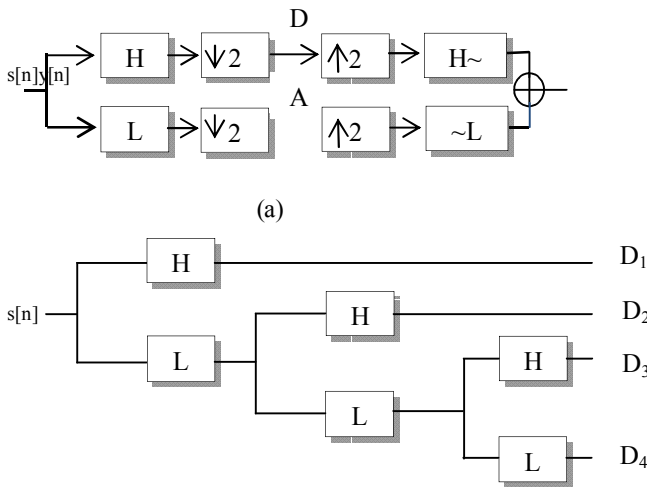


Fig.4. (a) Two-channels filter bank (b) 3-level octave band filter bank

4. FEATURE MATCHING

Feature Matching corresponds to study how iris can be identified. The matching between input image and image already Stored in database are matched. The feature vectors are matched in two images to conclude about the similarity and dissimilarity of the image.

The technique used for matching feature vectors is the Euclidean distance (Euclidean distance is the “ordinary distance” between two points that one would measure with a ruler and is given by the Pythagorean formula) based approach, unlike the traditional probabilistic neural network (PNN) approach using PSO, a new bio-inspired optimization method, as the learning algorithm. We calculate the Euclidean distance between the feature vectors and smaller is the distance, less is the dissimilarity between the two vectors.

Using various values of threshold, we calculate False Acceptance Rate (FAR) and False Rejection Rate (FRR) and hence get the Equal Error Rate (EER), as the intersection point of FAR and FRR curve. The high performance of iris verification system is in low Earthen testing images are matched with training set images. Using various values of threshold, we calculate False Acceptance Rate (FAR) and False Rejection Rate (FRR) and hence get the Equal Error Rate.

5. EXPERIMENTS & RESULTS

The entire algorithm was implemented using MATLAB 7.6.0.324(R2008a) .A set of iris images taken from 102

individuals was worked upon and tested. The training and testing of images was done. In order to understand the implementation of the algorithm we performed segmentation of iris images .Presently we have worked upon the segmentation in two and three parts. The study was done by using sobel operator and without using sobel operator. The results are as follows :

- 1) Results for two segments Image using sobel operator

Accuracy = 9.090909e+001=90.90909%
Equal Error Rate = 1.869345e+001=18.69345%
 FRR @ 1/100*FAR = 1.421195e-002
 FRR @ 1/1000*FAR= 4.737316e-003
 FAR @ 1/1000*FRR= 0

- 2) Results for two segments Image without sobel operator

Accuracy = 9.043062e+001= 90.43062%
Equal Error Rate = 2.011938e+001=20.11938%
 FRR @ 1/100*FAR = 1.894926e-002
 FRR @ 1/1000*FAR= 9.474632e-003
 FAR @ 1/100*FRR= 0
 FAR @ 1/1000*FRR= 0

- 3) Results for three segments Image using sobel operator

Accuracy =9.282297e+001=92.82297%
Equal Error Rate = 1.676536e+001=16.76536%
 FRR @ 1/100*FAR = 1.421195e-002
 FRR @ 1/1000*FAR= 4.737316e-003
 FAR @ 1/100*FRR= 0
 FAR @ 1/1000*FRR= 0

- 4) Results for three segments image without using sobel operator

Accuracy = 8.755981e+001=87.55981%
Equal Error Rate = 2.060732e+001=20.60732%
 FRR @ 1/100*FAR = 1.894926e-002
 FRR @ 1/1000*FAR= 9.474632e-003
 FAR @ 1/100*FRR= 0
 FAR @ 1/1000*FRR= 0

REFERENCES

- [1] J. Daugman, *Biometric Personal Identification System Based on Iris Analysis*, United States Patent, no. 5291560, 1994
- [2] J. Daugman, "High Confidence Visual Recognition of Person by a Test of Statistical Independence," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 15, no. 11, pp1148-1151, Nov. 1993
- [3] C. Sidney Burrus, Ramesh A. Gopinath, Hai Ta Guo, "Introduction to Wavelets and Wavelet Transforms: A Primer", 1998.
- [4] J. Daugman, "Demodulation by Complex-Valued Wavelets for Stochastic Pattern Recognition," *Int'l J. Wavelets, Multiresolution and Information Processing*, vol. 1, no. 1, pp. 1-17, 2003.
- [5] J. Daugman, "Statistical Richness of Visual Phase Information: Update on Recognition Persons by Iris Patterns," *Int'l J. Computer Vision*, vol. 45, no. 1, pp. 25-38 2001.
- [6] Li Ma, Tieniu Tan, Yunhong Wang, and Dexin Zhang, "Personal Identification Based on Iris Texture Analysis", *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 25, no. 12, pp. 1519-1532, Dec. 2003
- [7] R. Wildes, J. Asmuth, G. Green, S. Hsu, R. Kolczynski, J. Matey, and S. McBride, "A Machine-Vision System for Iris Recognition," *Machine Vision and Application*, vol. 9, pp. 1-8, 1996.
- [8] R. Wildes, "Iris Recognition: An Emerging Biometric Technology," *Proc. IEEE*, vol. 85, pp. 1348-1363, 1997.
- [9] Christel-Loic Tisse, Lionel Torres, Michel Robert, "Person Identification Based on Iris Patterns," *Proceedings of the 15th International Conference on Vision Interface*, 2002.
- [10] L. Ma, Tan, Tieniu, Wang, Yunhong, "Iris Recognition Using Circular Symmetric Filters", *Processing of THE 16th International Conference on Pattern Recognition*, vol.2, pp. 414-417, 2002.
- [11] Mayank Vatsa, Richa Singh, and P. Gupta, "Comparison of Iris Recognition Algorithms", *Proceedings of ICISIP'04, India*, pp.354-358, 2004.
- [12] Sanchez-Avila C., Sanchez-Reillo R.; de Martin-Roche D., "Iris Recognition for Biometric Identification Using Dyadic Wavelet Transform Zero-Crossing", *Proceedings of the IEEE 35th International Carnahan Conference on Security Technology*, pp. 272-277, 2002.
- [13] Ching-Han CHEN, Chia-Te CHU, "High Efficiency Iris Feature Extraction Based on 1-D Wavelet Transform", *2005 Design automation and test in Europe (DATE2005)*, Munich, Germany, March, 2005.
- [14] Jaideva C. Goswami and Andrew K. Chan "Fundamentals of Wavelets" 1999.
- [15] D.F. Specht, "Probabilistic Neural Network for Classification, Map, or Associative Memory", *Proceeding of the IEEE International Conference on Neural Network*, vol.1, pp525-532, 1988.
- [16] Chia-Te CHU and Ching-Han CHEN, "The Application of Face Authentication System for Internet Security Using Object-Oriented Technology", *Journal of Internet Technology*. (Accepted)
- [17] Ching-Han CHEN, Chia-Te CHU, "An High Efficiency Feature Extraction Based on Wavelet Transform for Speaker 2004 International Computer Symposium (ICS2004)", Taipei, Dec 2004.
- [18] Ching-Han CHEN, Chia-Te CHU, "Combining Multiple Features for High Performance Face Recognition System", *2004 International Computer Symposium (ICS2004)*, Taipei, Dec 2004.
- [19] Ching-Han CHEN and Chia-Te CHU, "Real-Time Face Recognition Using Wavelet Probabilistic Neural Network", *Journal of Imaging Science and Technology*. (Accepted)
- [20] J. Kennedy et al. "Particle Swarm Optimization", *Proc of IEEE Int. Conf. Neural Networks*, vol. IV, pp.1942-1948, 1995.
- [21] CASIA iris database. Institute of Automation, Chinese Academy of Sciences, [Online]. <http://www.sinobiometrics.com/casiairis.htm>.

Structuring of Normative-Technical Documents in Information and Computing Environment Taking into Account Ageing of Information

V.G. Lim¹, I.G. Voevodin¹, N.B. Tashpulatova², B.T. Kabulov³,

¹Astrakhan State University, Astrakhan, Russia

²Tashkent University of Information Technologies, Tashkent, Uzbekistan

³UzNeftegazinformatika, Tashkent, Uzbekistan

Abstract – In this paper the problem of uneven ageing of normative-technical information for the design, construction and operation of oil and gas pipelines is considered. The quantitative indicators for the document importance evaluation are proposed. The method of splitting the document library into separate warehouses and logical document removal (moving) is suggested.

Keywords – Quality Prognostication, Building Production, Industrial System, Repair and Construction Work, Analytic Planning, Analytic Hierarchy Process, Decision Making Support System, Data-Driven Application.

The developed information and retrieval system for the required normative documentation archive organization and information retrieval is a program product PLSystem / PLCode (Pipeline System / Pipeline Code) providing the database storage on CD or network server [1].

Under the scarce resource conditions the problem of the normative technical document corpus growth is apparently of the primary importance. As new documents must be continuously arranged, array sizes, and so, the document processing and storage cost will grow unlimitedly if do not take necessary steps for the corpus stabilization. The stable situation is attained when the intensity of normative document newly entering the corpus flow becomes equal to the removed document flow intensity. If corpus integration and document base completion procedures are sufficiently standardized, the out-of-date and not valuable material removal problem is insufficiently developed.

The out-of-date document removal necessity is stipulated by a number of interrelated factors including document corpus size augmentation, complication of really necessary document retrieval or document corpus management cost increase in the presence of budget limitations. The removal may consist of different operations, for example, of copying and removal of unnecessary materials or document displacement into the less expensive and less available warehouses. In the latter case a standard document warehouse (database) turn into a hierarchic multilevel warehouse system (distributed database).

The document corpus purge problem in an information and retrieval system is complicated by that the exact methods of material use accounting used in traditional libraries are not always sufficient.

So, some document types may require a special approach: laws acting at different periods, regulations and orders, documents of co-operating branches of production, foreign and extra-valuable materials, and documents on some subject areas in specialized editions [2].

Among the proposed quantitative indicators for the document importance evaluation the three ones are of particular interest: document age, i.e., number of years from its first edition; number of references to this document in posterior literature (for example, references number in documents published later); document use indicator determined by the number of calls to the database to search for the document text and entry.

All these indicators (besides the age) are hardly determined, and necessary parameters even if they may be determined usually have not a universal significance. The document removal decision, hence, in many cases is based on incomplete information. If we take into consideration, in addition to fundamental difficulties the cost of the removal operation itself linked to the necessity of numerous document corpus CD copies updating we can understand why effective procedures of corpus purge are so rare. But, as the removal has not good alternative (in view of the incessant document growth in all fields of knowledge) it would advisable to discuss some approaches to the problem.

The logical document removal in the information and retrieval system PLCode is executed by the displacement of their texts into a less available warehouse (for example, an archive CD), thereby a document library is being forming with separate warehouses in which documents being much in demand and up-to-date are available by usual call. In that way, the normative technical document corpus is partitioned in such a way that the mean use factor or mean negotiability would be far above in the actual document corpus than in a less available corpus part. More precisely, if N is total document amount in the database before the corpus purge it is advisable to select such a part of $x \cdot N$ ($x < 1$) documents for the storage in the actual document corpus that wouldn't be too large to provide an acceptable search time and yet would contain the greater part of current operating normative technical documents [3].

REFERENCES

- [1] V. G. Lim, V. D. Shapiro, V. I. Eristov and Ju. N. Argasov, "The software package for maintaining a database of normative documents for the design, construction and operation of oil and gas pipelines", *Main and field pipelines: design, construction, maintenance and repair*, Moscow: TsONiK GANG, 1997, № 4, pp. 6-18.
- [2] V. G. Lim and Yu. N. Klimov, "Patterns of ageing of normative-technical documents used in the PLCode retrieval system", *Methods of systems analysis and computer-aided design of investment and organizational and technological processes in construction*, Moscow: The section "Building" RIA Novosti, 2001, № 1, pp. 41-45.
- [3] V.G. Lim, Yu. N. Klimov and I.G. Voevodin, "Using patterns of ageing of information to improve the efficiency of information retrieval", *Proceedings of the All-Russian scientific conference "Scientific service in Internet*, Moscow: Moscow State University, 2001, pp. 72-74.

Labeled Protection of Project Tasks in Collaborative Designing of Software Intensive Systems

P. Sosnin

Ulyanovsk state technical university
32, Severny Venetc
Ulyanovsk, 432027, Russia

V. Maklaev

Ulyanovsk state technical university
32, Severny Venetc
Ulyanovsk, 432027, Russia

S. Zhukov

Ulyanovsk state technical university
32, Severny Venetc
Ulyanovsk, 432027, Russia

Abstract- The paper presents the system of means for the labelled protection of project tasks in designing the software intensive systems. The specificity of suggested means is defined by the uniform hierarchical representation of tasks and the team of designers the real time relations between units of which are being provided by programmed agents. Protection means are embedded to the toolkit WIQA (Working In Questions and Answer) supporting the work of designers at the conceptual stage.

I. INTRODUCTION

Any Collaborative Development Environment (CDE) supporting the creation of Software Intensive Systems (SIS) is a specialized system of SIS-type. Therefore, the practice of designing the SISs is being applied by designers of the CDE-systems. It extends also to designing the means of the informational safety which should be embedded to the created CDE [1].

The class of CDE-systems was defined in [2] where main features of typical CDE have been specified. In accordance of these features the Rational Unified Process [3] can be qualified as an example of CDEs. Therefore, this technology and corresponding toolkit can be used as the richest sources of more detailed requirements and their specifications as for developing the new CDE so for creating its informational security [4] and [5].

The analysis of the named informational sources indicates on the following important features which should be taken into account in designing the security of CDEs:

1. A life cycle of any created SIS can be presented by a system of project tasks $S(\{Z_i\})$ the real time solutions of which are provided by a team of designers $T(\{D_j\})$.
2. Means of managing the life cycle includes a dynamic system of appointments $S(\{A_k\})$ connecting any designer D_j with the definite subset of solving tasks $S_j(\{Z_{ji}\})$ in any current moment of time.
3. Any task of the life cycle is being solved by the responsible designer (only one) who can address for the help to other members of the team or to other persons (stakeholders) interested in the development of SIS. Usually stakeholders are specified as a specialized group which is included to the team of designers for fulfilling the definite actions in the designing (in solving the project tasks)
4. The system of appointments supports the use of a number of roles $\{R_m\}$ which can play by designers and stakeholders in their relations with tasks $\{Z_i\}$.

We have used the named features in the system of means for the labeled protection supporting the informational security in the real time work of designers with project tasks. The use of the uniform hierarchical representation of $S(\{Z_i\})$ and $T(\{D_j\})$ the real time relations between which are being provided by programmed agents defines the specificity of the suggested and developed means of security.

II. REPRESENTATION OF PROJECT TASKS AND TEAM OF DESIGNERS

The suggested means of the labeled protection are created for the workflows supported the conceptual designing of SISs with the help of the toolkit WIQA (C#, Microsoft.Net). These workflows and toolkit are presented in detail in [6]. In this paper we shall describe these means only schematically from the viewpoint of protecting the project tasks.

In WIQA the life cycle of the creating SIS is being modeled by the tasks tree each task of which exists in the form of its question-answer model (protocol of question-answer reasoning of the designer about solving the task). Such models are kept in the database of the WIQA-server and each of them can be accessible for designers in client workplaces as it shown in Fig. 1 and Fig 2.

In WIQA all interfaces are visualized in Russian therefore explanatory names are used for indicating interface areas and units accessible for designer. There are three types of such interactive units – tasks Z_i , questions Q_j and corresponding answers A_j .

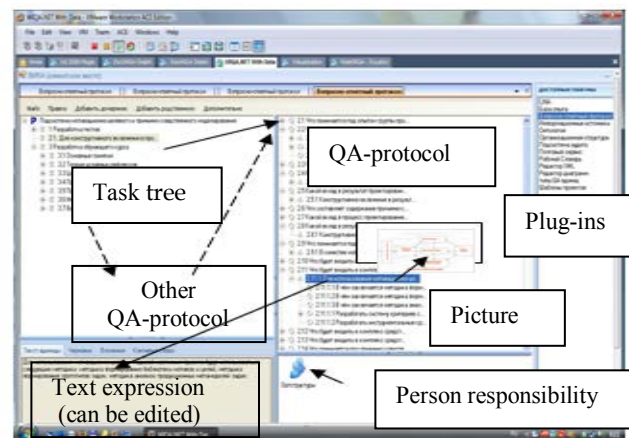


Fig. 1. Accessible state of project tasks (main interface)

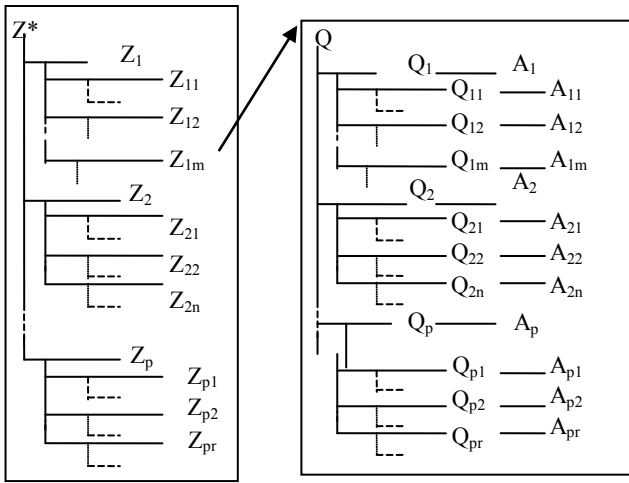


Fig. 2. Structure of data

Each of these units has a unique index name which creates automatically and includes the symbol of type (“Z”, “Q” or “A”) with concatenation of appointed indexes. The current state of the tasks tree and corresponding question-answer protocols (QA-protocols, QA-models of tasks) are dynamic constructions the typical view which is presented in Fig. 2. All of these interactive units are being produced (and are being evolved) by designers in the stepwise refinement process.

Relations between the described interactive models (the state of conceptual designing the SIS) are provided by special plug-ins named “Orgstructure” (organizational structure of the team of designers). For any task Z_i this plug-ins helps to appoint the responsible designer D_j and assistants $\{D'_k\}$ who can help in the decision process if in it there will be a necessity. In this case assistants will play the roles appointed to them.

Any appointed role defines not only a number of related skills and competencies but also responsibilities from viewpoint of the kind of the access to units of Z-, Q- and A-types (“the free access”, “only read the accessible unit”, “possibility to write the new unit” and “possibility to modify accessible unit”). Thus means of WIQA help to register the real time tasks structure of the creating SIS and relations of tasks with designers executing the appointed roles.

Plug-ins “Orgstructure” provides the realization of own functions with using of a definite part of the database named above. This part includes 24 relations defined by 155 attributes. Only subsets of these relations and attributes are used for registering the indicated appointments. Other functions of this plug-ins support the interacting between designers and real time managing in designing.

II. SUGGESTED APPROACH TO LABELED PROTECTION OF PROJECT TASKS

The described materialization of project tasks and appointments didn't concern the questions connected with checking the rights of the real time access to interactive units. Such checking is a task of the informational security which should be solved on the base of modern practices accumulated in corresponding international standards, first of all in the standard ISO/MEC 15408 [7]. The task of checking the rights has been formulated in the statement which correlates with this standard in its recommendations for the labeled protection profile [7]. The solution of the task has led to the subsystem of labeled protections included to WIQA and presented in Fig. 3 where interfaces are marked with explanatory labels (because interfaces in Russian).

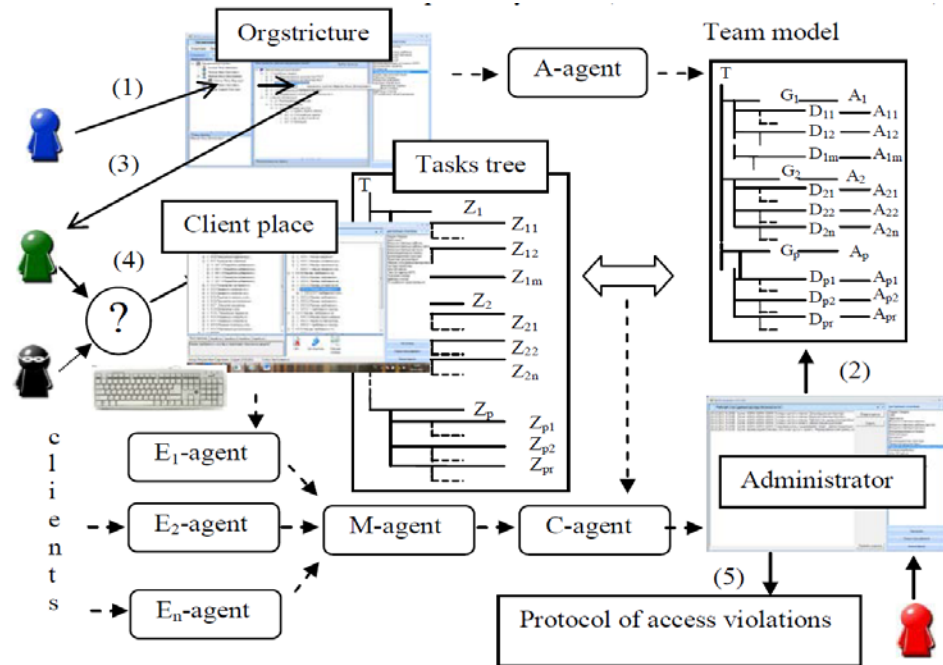


Fig. 3. General scheme of labeled protections

Specificity of the developed subsystem is defined by following features and actions:

1. The protection task is being solved with using the tasks tree and the copy of the appointments from the organizational structure. Solving this task is being implemented by the administrator who appoints labels of protection and manages the events of the access to the tasks tree. The administrator is worked on the specialized client place of WIQA.

2. The copy of appointments is loaded to data structures types of which are used for the tasks tree and QA-models of tasks (Fig. 1).

3. The real time copy of appointments is created and supported by the program agent named “agent of appointments” (A-agent).

4. Any access attempt of any designer to any task of the tasks tree is being fixed by the program agent acting at the corresponding client place. Such agent (named “agent of events”, E-agent) forms the flow of events each of which registers the designer and the name of the task chosen in the access attempt. Named events will be denoted below as A-events.

5. The program agent (named “agent of monitoring, M-agent” and acting at the WIQA-server) combines flows of events from all client places and extracts from flows the events which should be checked on the compliance of rights.

6. The program agent (named “agent of checking”, C-agent, acting on the workplace of security administrator) checks the rights of access attempts included to the entering flow. In accordance with results of tests the agent gives or permissions for access attempts or includes to the list of access violation new records.

The described features and actions are presented without details in Fig. 2

The scheme of protections indicates four typical versions of designer participations in security processes. The chief of designers group has right to choose (1) the responsible designer or assistant for the new task or problematic task. Any new appointment is revealed by the A-agent which is responsible for the adequate copy of all appointments in the team model correlating with the tasks tree.

Any new appointment is being processed (2) by the administrator who enciphers the relation between the corresponding units in the tasks tree (task Zi) and team model (new appointment).

After named appointments (1) and (2) the chosen designer (3) can activate the access (4) to the named task (Zi) but it can be or in accordance with appointed rights or with their violation. Let us notice that the access to the task Zi can be activated accidentally or intentionally by the designer who has not any right of the access to Zi.

Therefore, any attempt (3) of the access to the task Zi should be revealed. For such revealing the touchable actions of the designer (for example, actions via keyboard) are being processed by the corresponding agent of the E-type which sends the records about the potentially dangerous events to the M-agent. As told above the C-agent processes the flow of

events from the M-agent. Results of processing are used for managing the labeled protection of the project tasks.

III. AGENT OF APPOINTMENTS

The decision to use the copy of the team representation is explained first of all the necessity to separate the information about the team and appointments which uses by designers from the information about it which should use by the security administrator. Such division supports by the A-agent which extracts the necessary information from the subsystem of database (modeling the Orgstructure) and loads the team model in the subsystem of database which supports the work with QA-models (Fig. 4).

In duplication the A-agent provides the adequateness of descriptions as for the team so for appointments in both subsystems. The additional team model inherits all possibilities which the tasks tree and QA-models presents for designers but such possibilities in interactions with the team model will be accessible for the administrator only.

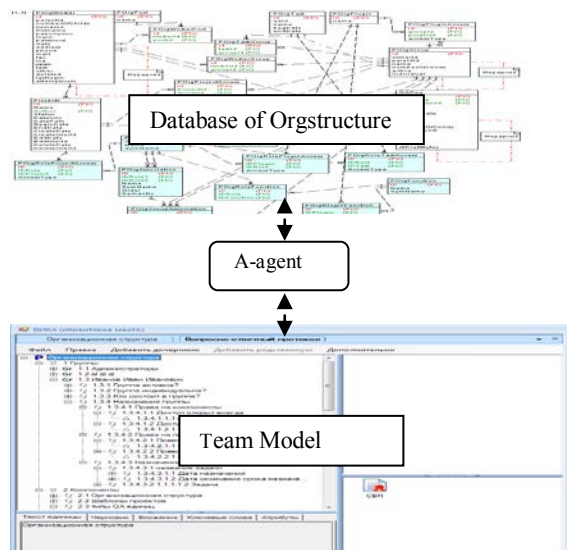


Fig. 4. Duplication of information about the team and appointments

In duplication the A-agent provides the adequateness of descriptions as for the team so for appointments in both subsystems. The additional team model inherits all possibilities which the tasks tree and QA-models presents for designers but such possibilities in interactions with the team model will be accessible for the administrator only.

IV. CRYPTOGRAPHIC PROTECTION WITH LABELS

The inclusion of labels into relations of the protection between members of the team and project tasks is the main function of the security administrator the work of whom is being managed by new appointments of designers in solving project tasks [8]. The content of this function is generally presented in Fig. 5.

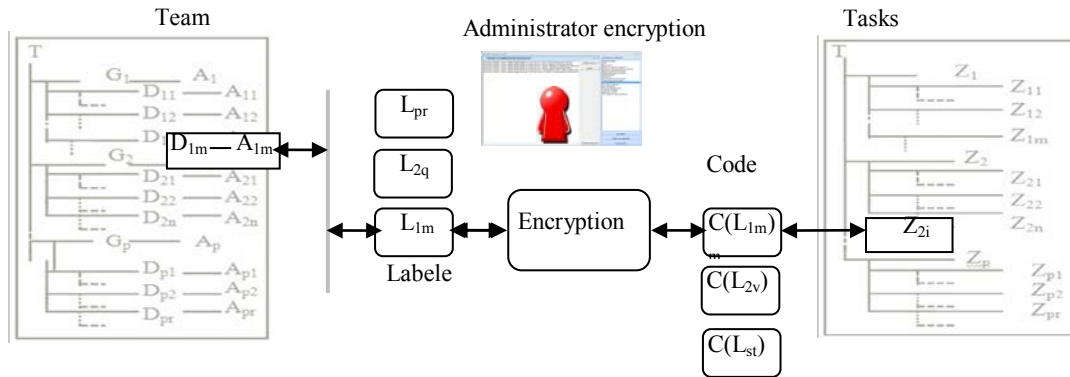


Fig. 5. Place of labels in protection of tasks

Let us assume that task Z_{2i} is appointed with right A_{1m} to designer D_{1m} and this appointment $A_{1m}(D_{1m}, Z_{2i})$ loads to the team model. Such loading is the event which requires (from the administrator) to bind by the label the appointment $A_{1m}(D_{1m}, Z_{2i})$ with the task Z_{2i} in the tasks tree. For this the administrator generates the new label and adds its name to the list of labels each of which binds the designer D_{1m} with another task appointed earlier. This list $R(\{L_{mn}\})$ is included to the team model. Any label in this list is enciphered and included to the list of codes each of which binds the task Z_{2i} with the corresponding appointments for this task. The list of codes $R(\{C(L_{pq})\})$ is attached to the task as the additional attribute.

Information in both these lists is used for the real time checking of access attempts to the task Z_{2i} . In any time the

administrator can change the encryption procedure on another procedure [9] which is being extracted from the specialized library.

V. AGENTS OF EVENTS AND MONITORING

The second type of information for solving the task of labeled protection is “T-event” registering the access attempt to the project task. In the suggested approach such events are being reviled in the protocol of touchable actions of the designer, which is registered in the corresponding client place. In Fig. 6 the keyboard fulfils the role of the source of key-driven operations.

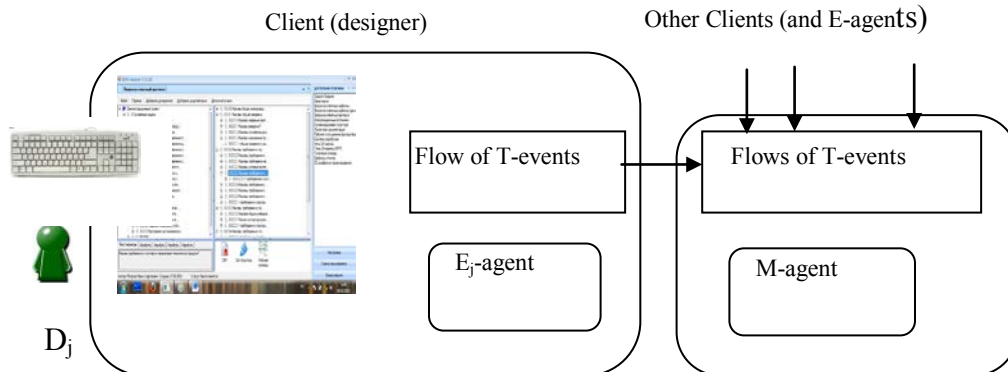


Fig. 6. Forming flows of A-events

All visual components of the main interface of WIQA are successors of class Windows. Forms. Control, including the main form. It allows to use interface Control for the reference to any object of the form, for discovering the events indicating the interaction of the designer with the chosen task.

After activating the definite keyboard operation km , the E_j -agent creates and includes the current unit A_{jk} to the flow of T-events $\{A_{jk}(t_n, D_j, Z_i, S_r, T_q)\}$, where: t_n - time of occurrence of T-event, D_j - the designer from whom there was an event, Z_i

- name of the chosen task, S_r – tasks subordinated to Z_i , T_q - type of action with Z_i . The M-agent combines messages from a number of E-agents in flows of T-events on the WIQA-server.

VI. AGENT OF CHECKING

Any A-event from their flow in the server is being processed by the C-agent which is responsible for timely checking the

right for any access attempt. The scheme of checking is generally presented in Fig. 7

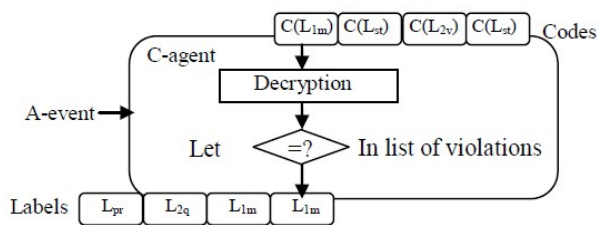


Fig. 7. Processing of A-events

The task name Z_j from description A_{jk} of the access attempt is used by the C-agent for extracting the list of codes $R(\{C(L_{2pq})\})$ for this task from the tasks tree. Elements of this list are being decrypted and combined in the list $R(\{L_{3pq}\})$. Similar way but with using the designer name D_j the C-agent chooses the list $R(\{L_{1mn}\})$ from the team model. Comparing lists $R(\{L_{1mn}\})$ and $R(\{L_{3pq}\})$ gives an answer about the right of the access attempt. The positive result opens the access to the designer while negative results are being registered in the protocol of violations.

VII. ADDITIONAL APPLICATIONS

Nowadays the presented means are being adjusted for their use in two applications. One of these applications is the labeled protection of assets in the experience factory (EF) of the project organization.

The experience factory has a catalogue of assets the structure of which can be interpreted similarly to the tasks tree. But in this case (Fig. 8) the identifiers of catalogue sections and subsections are used instead of identifiers of project tasks.

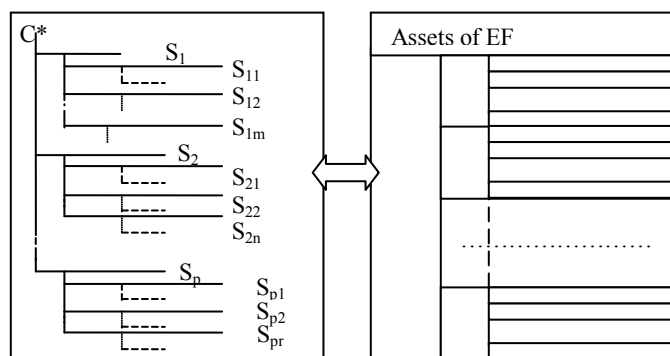


Fig. 8. Structures of catalogue and assets

All told above about the labeled protection of tasks is right for the protection of sections and subsections of the catalogue. The difference is in the structure of the corporate network. The experience factory supports the processes in the network combining a number of deployments of WIQA.

The second application is a corporate e-mail to be implemented by means of WIQA. In this case informational flows are organized in frames of corresponding tasks the labeled protection of which was presented above.

VIII. CONCLUSION

The described means of the labeled protection have confirmed the stability of functioning in their use in the real practice of designing. Following directions of their applications are developed and planned: the adaptation of means for the work of designers with the repository catalog providing the access to units of the experience factory; the labeled protection of the corporate e-mail; full informational separating the means of protections from the database of designing (duplicating the models of the team and tasks tree into the database of the labeled protection).

REFERENCES

- [1] M.V. Stringfellow, N.G. Leveson and B.D. Owens. "Safety-Driven Design for Software-Intensive Aerospace and Automotive Systems". Proceedings of the IEEE vol. 98, Issue 4, pp. 515-525, 2010.
- [2] G. Booch and A. W. Brown, "Collaborative Development Environments," [http:// www. booch. com/architecture/blog/artifacts/CDE.pdf](http://www.booch.com/architecture/blog/artifacts/CDE.pdf), 2004.
- [3] P. Kroll and Ph. Kruchten, "The Rational Unified Process Made Easy: A Practitioners Guide to the RUP," Addison-Wesley, 2003.
- [4] B. D. Owens, M. S. Herring, N. Dulac, N. G Leveson, M. D. Ingham and K. A. Weiss, "Application of a safety-driven design methodology to an outer planet exploration mission". In *Proc. IEEE Aerospace Conference*, pp. 1-24, 2008.
- [5] G. H. Ramirez Caceres and Y. Teshigawara, "Study on a Threat-Countermeasure Model Based on International Standard Information". The 12th World Multi-Conference on Systemics, Cybernetics and Informatics (WMSCI 2008), Orlando, Florida, USA. pp.227-232, 2008.
- [6] P. Sosnin, "Means of question-answer interaction for collaborative development activity". Hindawi Publishing Corporation, *Advances in Human-Computer Interaction*, Volume 2009, 18 pages, 2009.
- [7] ISO/IEC 15408 "Common Criteria for Information Technology Security Evaluation Part 1~3 Version 3.0", June 2005.
- [8] L. Zheng and A. C. Myers "Dynamic Security Labels and Static Information Flow Control". *International Journal of Information Security* Volume 6, Issue 2, pp. 67-84, March 2007
- [9] C. Liu, A. Billard, M. Ozols and N. Jeremic, "Access Control Models and Security Labelling". In *Proc. The Thirtieth Australasian Computer Science Conference*, pp 181-190, 2007.

Integrated Analytical Information Resource Management System

Giorgi Ghlonti

Dept. of Informatics and Programming
N. Muskhelishvili Institute of Computational Mathematics of the Georgian Technical University
Faculty of Information Technologies and Engineering
International Black Sea University
Tbilisi, Georgia
gg59ster@gmail.com

Abstract— The paper is considered to the problems of development of analytical information resource management systems. The authors present a service-oriented architecture solution that provides data collection and aggregation at the points where information emerges. The user is provided by a set of functional complexes, allowing to build the applications covering the entire lifecycle of analytical information resource from the planning data collection to the stages of data processing.

Keywords— analytical information resource management; service-oriented architecture; functional complexes; data warehouse; homogeneous information model

INTRODUCTION

The accumulation and effective use of information, describing the status and activities of any social and economic system appears to be the condition of its successful development.

High quality information resource is an asset that determines the efficiency of decision making at all levels of hierarchy in social and economic environments.

Effective use of this asset requires the development of strategy for managing the analytical information resource, which implies not only the regulation of local problems concerned with data treatment at the points where information is gathered or aggregated, but also the decisions concerning the global issues of unity, completeness and consistency of analytical information resource throughout the subject area, integrity and transparency of information space.

In this paper the author present a service-oriented architecture solution that provides data collection and aggregation in the points where information is gathered. The entire lifecycle of analytical information resource, from the planning data collection, to the stages of processing and analyzing the results is covered.

System requirements include:

- Adaptability to the diversity and multiple criteria of subject area;
 - Support of open and flexible information model;
 - Ability to locate significantly large amounts of data in common information space, where they would be used by all software applications that have access to data warehouse;
 - Guarantee completeness, consistency and integrity of information resource;
 - Provide users by tools of building application software without programmers intervention;
 - Minimize time and cost for data processing;
 - Provide homogeneity of information model, in order to ensure transparency of the information space.
- The project includes following components:
- The subsystem for collecting, storing and multiple reuse of information;
 - Functional complexes for data processing;
 - Technology for designing user applications.

DATA WAREHOUSE

In the design of the subsystem for collecting, storing and multiple reuse of information the authors were guided by considerations of analytical information to be, in essence, statistical data and that their collection and processing is subject to the requirements and laws that regulate the statistical studies and observations in general.

Thus the system requirements should be interpreted as the need to provide the user with a mechanism that allows the planning and conduct of various statistical studies, placing the results in the common information space, with the ability to use these data as an analytical source of information by all applications and systems from subject area.

It should be take into account, that the accumulation of knowledge in the subject area requires the development of some general model of it, according to which the researches are planned. In this case, data will be accumulated gradually

and above mentioned quality standards for information resource and information space will be provided.

The organizational aspect of data collection for statistical researches is presented in a statistical document.

The hierarchical structure of the document corresponds to the organization of information gathering process. The document represents the model of the environment in the form of a set of parameters that make up the surveillance program, along with the set of possible values for each of them. This set is classified according to the unity of time and place of observation.

In the case of formalizing a document an analytical model is created that allows variety of processing techniques to be applied to the information collected on its basis.

The formalization of the document is also important because its structure makes possible to plan the structure of information space, where data are placed.

In [1] a formalized model of the document is presented in the notation of UML. In this model the textual part of a document is separated from its geometrical structure, and is described separately from it.

This means that the same document may be stored in different languages, while the data, collected on its basis will be stored in a single copy.

In the given model of document, each component of it, including lowest one – the sell, receives a classification code, called a Coordinate Index (CI). CI clearly defines the position of given component of document and can be used to refer to an element of data afterwards. For this reason the data warehouse is supported by a hierarchy of indices [2]. The data model obtained, being homogeneous, has advantages before known structures usually used to build dimensional models.

The subsystem for collecting, storing and reusing information includes the following functions:

- The construction of the model of a subject area and its representation in the form of an electronic document;
- Collection of information with possibility of its multiple reuse (data warehouse);
- Some tools for information retrieve.

FUNCTIONAL COMPLEXES

Functional complexes are service components, organized as independent software units. By means of them the user has ability to request execution of various functions of data processing. Among the functional complexes we mention:

- Interpretation of algebraic and logical expressions;
- Solving of algebraic equations;
- Ordering according given hierarchy of parameters;

- Identification on the basis of the cipher of unit of observation;
- Analysis of variance.

Functional complexes are managed by so called orders. Those are regular expressions, composed on the basis of certain grammar [3]. CIs of sells are the operands of those constructions.

TECHNOLOGY FOR DESIGNING USER APPLICATIONS

Technology for designing user applications provides the user with a mechanism of building applications on the basis of functional complexes. The user has the opportunity to write a script for applications using appropriate functional complexes. There is also a Scheduler, providing optimal sequence of orders for each functional complex.

Among the simplest user applications the calculation of indicators, monitoring the consistency of the data, statistical analysis, problems of operations research, the study of cause-effect relationships – data mining, etc. may be mentioned.

With their help you can build a variety of data processing systems. A wide range of opportunities is covered, including both the so-called routine data processing tasks (e-government, financial accounting, managerial accounting, etc.) as well as the problems related to the traditional functions of management information system, strategic planning, virtual modeling, etc.

CONCLUSION

Thus we have a service-oriented solution providing the user by tools for building different data processing systems on different level of social and economical hierarchy. The entire lifecycle of information resource is covered and user may construct information space where this resource would be gradually gathered and used by all stakeholders as global asset and the condition for sustainable development.

REFERENCES

- [1] A Chaduneli, M. Pkhovelishvili, G. Ghlonti. "The model of electronic document used in the organization of statistical surveys. Transactions of the Georgian Technical University, vol. 3, pp. 30-35, Tbilisi, 2004. A. Чадунели, М. Пхovelishvili, Г. Глонти. "Модель электронного документа, используемого при организации статистических наблюдений", т. 3, стр. 30-35, Тбилиси, 2004.
- [2] I. Ghlonti, T. Maruashvili, K. Meskh. "On the method of storing the documents in computer". Proceedings of N. Muskhelishvili Institute of Computational Mathematics of Georgian SSR Academy of Sciences, pp. 45-50, Tbilisi, 1986. И. Глонти, Т. Маруашвили, К. Месхи. "Об одном методе хранения документов в памяти ЭВМ", Труды Института вычислительной математики им. Н. Мухелишвили АН ГССР, стр. 45-50, Тбилиси, 1986.
- [3] G. Ghlonti. "Problems of construction of unified analytical information space for organizational management support"; PhD thesis, Tbilisi, 2006. Г. Глонти. "Проблемы построения единого информационного пространства управления организацией" диссертация на соискание степени кандидата технических наук. Тбилиси, 2006.

NFC : Smart Recording Of Traffic Violation System

Omid Nejati

Member Of Young Researchers Club
Islamic Azad University Qaemshahr Branch
Qaemshahr, Iran
Omidnejati.it@gmail.com

Mohsen Yaghoubi Suraki

Department of Computer Engineering
Islamic Azad University, Qazvin Branch
Qazvin, Iran
m.yaghoubi@qiau.ac.ir

Abstract— Traffic violations leading to car crashes results in thousands of deaths each year in the world, most of which occur because of breaking speed limits and unauthorized overtaking. Traffic experts believe that management of vehicles on dangerous roads is considered to be one of the main factors in reducing car crashes. In this regard using electronic equipment can lead to a reduction in roadway crashes. In this article we are going to introduce a smart system designed via RFID and NFC technology. This system can intelligently register drivers' violations including unauthorized speed and overtaking. The system also furnishes other technologies including speed and line tracking sensors. The recorded violations will be immediately dispatched to the central server of traffic police via driver's cell phone NFC enable, the system is implementing with RFID technology. Finally a fine will be levied on driving offender and it provides opportunity for smart development of transportation and electronic traffic police. Therefore, this cycle forms a smart violation recording system. By means of the system, drivers will be under police control in each time and there will be no conspiracy between police officers and violator. You can also reduce human factors involved in car crashes and increase supervision on law enforcement.

Keywords-RFID; NFC; Traffic Violations; Traffic Police.

I. INTRODUCTION

Controlling traffic laws observance has become a tough task because of growth of cities and roads and mass production of cars in all countries of the world including Iran. By means of novel technologies, the number of human factors involved in car crashes decreases and consequently drivers and roads will be under close supervision. One of these novel technologies is called NFC.

Using this technology with two specific traffic signs, we can design a system which is able to control speed limit breaking in addition to unauthorized overtaking. We can also improve safety in cities and roads and minimize the number of casualties resulting from these accidents. So far similar projects have only concentrated on vehicles' speed; however, in this article we are going to introduce you a system which is able to establish control over drivers who break speed limits and do unauthorized overtaking. In this system, traffic violations are recorded and then the information will be sent to traffic police servers via drivers' cell phones. In this paper we use NFC technology for data transmission because NFC

technology by default used in many mobile device in the world. First we discuss about RFID and NFC technologies and then explain about the way you can mount smart system in the mentioned traffic signs and we discuss about violation recording and transmitting system. Finally we enumerate the system's advantages.

II. RFID

First RFID means using radio signals in order to identify an object automatically on the basis of distant data storage and retrieval [1].

Basically, implementing RFID technology needs the following equipment:

- Tags
- Tag Reader
- Antenna- Signal Booster
- Information Management Software
- Data Bank

Radio frequencies exchange information between data transmitter and receiver. The component which transmits information is called tag and the receiving part is called tag reader. Another way of classifying tags also exists. In this way tags are classified in terms of their energy sources and fall into three distinctive categories: active tags, passive tags and semi-passive tags. There are many differences between active and passive tags but the main difference between them is that active tags take the required energy from their accompanied battery whereas passive tags do not have energy source per se and must use the electromagnetic energy disseminated from tag reader in order to operate. Of course it has to be added that compared to active tags, passive ones have a limited reading domain [1]. Moreover passive tags have a lower price, longer lifetime and small size. Semi-passive tags are the other types of tags which can use tag readers' disseminated frequencies' energy in addition to its internal battery. In this article we will focus on passive tags.

The antenna is used to transmit radio signals between tag and tag reader which is used for both of them. The information management software is used for processing the collected data.

By means of this particular software which is normally implemented on a local server, data transmission and collection via tag reader will be possible. If it is needed, it can

also store and retrieve the aforementioned data in the data bank.

RFID technology can be a substitute for bar codes. Indeed this technology is beyond bar codes because it is equipped with an automatic scanner system. In other words there are dramatic differences between these technologies two of which is that RFID can store a large amount of information and it has no need to sightline in data collection and communication [2].

III. NFC

Near Field Communication is a short range wireless communication technology which is evolved from Radio Frequency Identification (RFID). It enables communication between two NFC enabled devices within few centimeters. Recently many business scenarios are implemented using NFC technology such as payment, ticketing, supply chain management systems, and also new NFC specific scenarios are arisen such as smart posters. NFC's three different operating modes, which are described later, and increasing capabilities of mobile devices enabled many NFC scenarios to be implemented.

In NFC, the communication occurs when two NFC compatible devices are brought together less than four centimeters, or simply by touching themselves. It operates at 13.56 MHz and can transfer data up to 424 Kbits per second [3]. In an NFC model two devices are involved in the communication, which are called initiator and target. Initiator is an active NFC device which is responsible for starting the communication. Also it has an embedded energy component whereas target can be either a tag, RFID card or an NFC device which responds the initiator's requests [4].

Other most important advantages of NFC technology include:

- The technology is compatible with existing RFID structures, existing RFID tags and con-tactless smart cards [3].
- It is easy to use and familiar to people because users don't need to have any knowledge about the technology. All a user has to do is to start communication by bringing two devices Together [5].
- The transmission range is so short that, when the user separates two devices, the communication is cut. This brings inherent security. If there isn't any other device close, there is no other communication.

NFC has three operating modes; Peer-to-Peer, Reader/ Writer, and Card Emulation. These operating modes are defined by NFC forum [3], which was formed to advance and standardize the use of Near Field Communication technology.

In card-emulation mode the data is transferred from mobile device to NFC-Reader; in reader/writer mode data is transferred is from NFC tag to mobile device or mobile device to NFC tag; and in peer-to-peer mode data is transferred between two NFC compatible devices [5].

IV. IMPLEMENTING RFID TECHNOLOGY IN VIOLATION RECORDING SYSTEM

As we discuss in the previous section, implementing RFID technology needs tags to store traffic signs information . Because of enormous number of traffic signs, allocating

separate tags to each sign is impossible so in this article we will focus on two specific signs, ignorance to which causes lots of car crashes each year:

- No overtaking signs
- Speed limit signs

There are a large number of these signs in cities and roads, so in order to bring down costs you can use passive tags with only reading memory (because the information within these tags will never changed), moreover these tags have a longer lifetime and do not need battery.

The aforementioned signs have to be installed hundreds of meters far from the specific event, when a car approaches the sign i.e. a few meters distant from it, emitted magnetic frequencies from tag readers which have been implemented in violation recording system within the car will activate the sign's tag and then its in-formation will be dispatched to the tag reader then the information within tag's memory will be read (Fig. 2).

Within the system there is very simple software which has been designed in order to adapt the information with sensors and in case of occurrence of any mistakes the mentioned software will transmit them to the drivers' cell phone via radio frequency receive or transmitter [6].

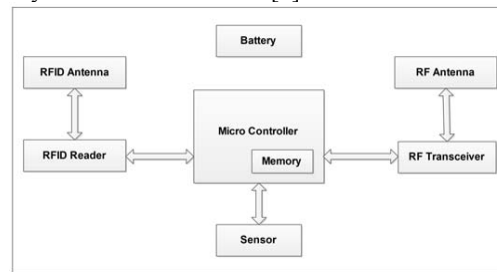


Figure 1. Violation Recording System in Cars[7]

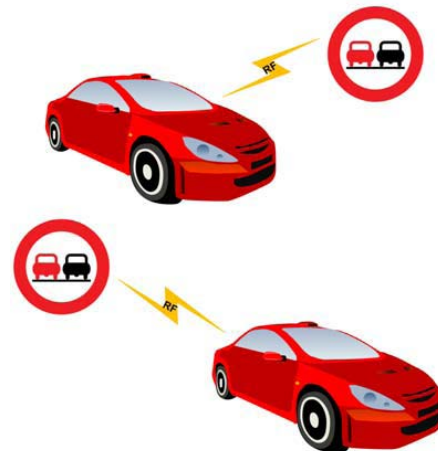


Figure 2. Different Stages Of Ttraffic Identification Via RFID Technology

The following is how you can make the technology feasible in the aforementioned traffic signs.

A. No Overtraking Sign

This sign mostly is used in roads with more turns and insufficient sight distance. Latest statistics prepared by traffic police indicates that in Iran at 2008, 25% of roadway crashes occurred because of unauthorized overtaking [9].

Violation recording system provides the opportunity to establish control over vehicles' overtaking in the no overtaking zone automatically. First, no overtaking sign is identified by the system and in case of any unauthorized overtaking by drivers the traffic violation will be recorded. In order to make the system feasible you can utilize line tracking sensors.

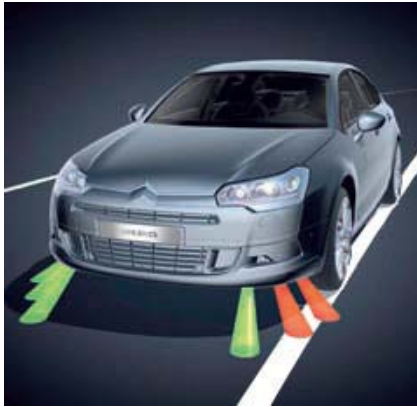


Figure 3. Utilize line tracking sensors [8]

As you can see in(Fig. 3) when the left sensors activate, driver receives warning and when the right sensors become activated the violation will be rec-orded. It has to be noted that no overtaking zone have to be specified with two signs, one of which shows the beginning of limits and the other indicates the end of no overtaking zone(Fig. 4). At the end the system will return to its first mode [6].



Figure 4. Specifying no overtaking

B. Speed Limit Signs

These signs limit speed for drivers in specific zones, so drivers can have control over their vehicles in confrontation with unexpected accidents. According to the statistics provided by traffic police, breaking speed limits have caused 20% of car crashes in Iran at 2008 [9].

Traffic experts believe that management of vehicles on dangerous roads is considered to be one of the main factors in reducing car crashes. In this regard using electronic equipment can lead to a reduction in roadway crashes [10].

By means of violation recording system, you can control vehicles' speed automatically. At the beginning of speed limit zone drivers have to be informed of authorized speed. Then the authorized speed will be stored in the system's memory. Whenever drivers break the speed limit, they will receive a warning and in case of ignorance to the warning the traffic violation will be recorded. The system is connected to the vehicles' speed sensors and controls its speed in each moment. It has to be noted that the beginning and the end of speed limit zone have to be specified by two traffic signs. Leaving the aforementioned zone, the system will return to its first mode [6].

V. TRANSMISSION SYSTEM OF TRAFFIC VIOLATIONS

This system constitutes the last stage of violation recording cycle in traffic department. In this system the recorded violations are dispatched to police central server in real time. To accomplish the goal we will use the structure shown in (Fig. 1) [7].The cell phone itself is equipped with a NFC. Cell phone received information by NFC from RF transceiver. Then the cell phone transmits the information using GSM and GPRS or Wi-Fi to the central server (Fig. 5). So by means of aforementioned technology direct communication will be possible. At this time the number of license plate of vehicle will be sent for the server together with the recorded traffic violations. Since the information is sent in real time the time, date and location (by GPS) of violation are also recorded.

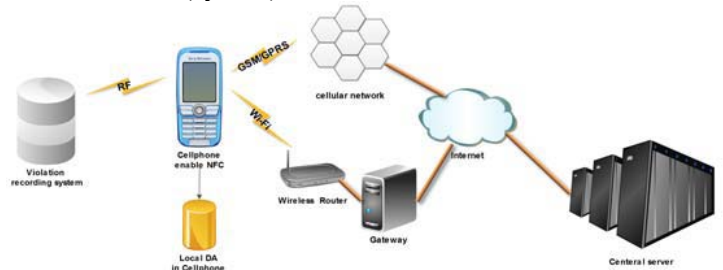


Figure 5. Transmission system

VI. CONCLUSION

Using RFID technology together with two traffic signs and NFC technology we designed a system which is able to receive the information within the sign's tag and register the information for the violator, the system has the following advantages:

- Minimizing the possibility of violator's escape (guilty driver) through conspiracy or when police officer is absent or busy.
- Human forces are substituted for mechanized equipment and consequently eliminating human errors including visual error, fatigue, etc. will be eliminated. The number of required trained human forces will be de-created and as a result you can bring down extra costs.

Manipulation or changing vehicle's license plate to prevent identification will be useless.

Easy identification of public transportation vehicles which break speed limits will be possible; as a result safety in public trips will be improved.

In case of car heist or a vehicle being under prosecution, the system will have a significant contribution in its detection or stopping.

Some of traffic violations are identified in a mechanized way so filing their information will be an easy task and there is no need to register them manually.

This system can especially be used in public transportation system of our country and also to reduce the number of human forces in traffic department and to establish control over all of the country's roads and streets.

Coverage of other violations such as stopping at a no stopping zone, not leaving enough space between two vehicles, and control establishment over restricted traffic areas will be possible by integrating other capabilities into this system.

REFERENCES

- [1] Fine, C., Klym, N., Trossen, D., and Tavshikar, M., The Evolution of RFID Networks: The Potential for Disruptive Innovation, MIT Center for eBusiness *CeB Research Brief*, Vol. VIII, No. 1, 2006.
- [2] Holmqvist, M., and Stefansson, G., Mobile RFID - A Case from Volvo on Innovation in SCM, System Sciences, *HICSS '06*, Proceedings of the 39th Annual Hawaii International Conference on, vol 6, IEEE, 2006.
- [3] NFC-Forum, Available: <http://www.nfc-forum.org>.
- [4] Ecma International, "Near Field Communication - White Paper", 2005, Ecma/TC32-TG19/2005/012, Available: <http://www.ecmainternational.org/>.
- [5] Kerem OK, Vedat COSKUN, Mehmet N. AYDIN, and Busra OZDENIZCI, "Current Benefits and Future Directions of NFC Services", International Conference on Education and Management Technology, Cairo, Egypt, November 2, 2010.
- [6] Nejati, O., "Smart Recording of Traffic Violations Via M-RFID" 7th international conference of Wireless Communication, Networking and Mobile Computing, Wuhan, China, September 23th, 2011.
- [7] Bahreininejad, A., Salman TAHERIZADE, "A Solution for Objects' Internet", 2nd International Conference of RFID, Tehran February 20st, 2008.
- [8] citroen-c5 <http://www.citroen.co.uk/new-cars/citroen-c5-saloon/>.
- [9] Saipa <http://www.saipaonline.com/view-1498.html/>.
- [10] News. "<http://www.fardanews.com/fa/pages/?cid=90965/>"

Rule Scheduling Methods in Active Database Systems: A Brief Survey

Hamid-Reza Firoozy-Najafabadi
Department of Computer Engineering
Science and Research Branch,
Islamic Azad University
Tabriz, Iran
hr.firoozy@iauotash.ac.ir

Ahmad Habibizad Navin
Department of Computer Engineering
Science and Research Branch,
Islamic Azad University
Tabriz, Iran
habibi@iauotash.ac.ir

Abstract— Traditional database management systems (DBMS) have static nature. This means that the operations such as query, update, etc. is performed only when asked by the user and DBMS doesn't act initiative in the specific condition. But many applications need to have automated monitoring, till in the occurrence specific event database management system will perform suitable reaction automatically. For this reason, active database systems (ADBS) were designed. One of the key issues in the active database management system is rules scheduling. This paper first introduces active database system and then will be investigated different methods of rules scheduling in the ADBS. Finally, scheduling methods is evaluated using active database system simulator (ADSS).

Keywords- active database system; rules scheduling; active database system simulator;

I. INTRODUCTION

Database is considered as set of integrated stored data (about types of Entities of an operation environment and their relationships based on defined formal structure, with minimum redundancy under centralized control that is used by one or more users simultaneously [1, 17]. Traditional database management systems should perform tasks such as maintenance, updating, inserting, deleting, reporting, etc. These DBMSs are static, Means that the operations such as query, update, etc is performed only when asked by the user and DBMS doesn't have any initiative in the specific condition. But many applications such as warehousing, factories automation and complex financial calculations systems (e.g. stock exchange) need to have automated monitoring, till in the occurrence specific event, database management system will perform suitable reaction automatically. For this reason, a database system should be designed to allow define events and their corresponding reactions. Such a system called active database system (ADBS) [1, 2].

Active database systems have been combined database technology with techniques such as rule based programming and event oriented programming. These systems supports rule definition, rule management and rule execution. In the other word, supports the execution of specific active behavior. For supporting reactive behavior of active database systems, Event-

Condition-Action (ECA) rule format was created. The ECA format has three sections: Event, Condition, and Action. When an event occurs, condition gets evaluated and if the condition is true, action is executed [9].

Work flow management systems, network management, financial systems, transportation systems financial calculations systems, etc have total or part of active behavior characteristics and are considered as suitable systems for using active database. Up to now, a lot of researches have been performed about integration and using dynamism techniques in relational and object oriented database management systems [1, 2].

Despite the advantages of using active mechanisms, but a few application have benefited from this feature. The reason is that users aren't sure about reliability and safety of active mechanisms. Of course, trust can be obtained through gain experience and skills in design and implementation of ADBS. Some of these problems include [1, 15, 16]:

- Application control flow is hidden from the viewpoint of users
- Rules definition and explanation is difficult
- Lack of standard and comprehensive language to define the rules

Of course, currently researchers have offered solutions for each of these problems.

After this introduction, paper is organized as follows: Section 2 offers an overview of active database systems and rule processing in the ADBS. In Section 3 we analyze existing rule scheduling methods in ADBS. Section 4 will provide a comparison and evaluation of existing rule scheduling methods.

II. ACTIVE DATABASE SYSTEM

A. Introduction to Active Database System

Active database systems are organized in the form of active rules set. These rules determine different states of the database and supports reactive behavior of active database systems. These rules include three parts that are: event, condition and

action. For this reason, they are called event-condition-action rules or briefly ECA. When specific event is occurred in the body of rule, condition part evaluated and if its value is correct, then action part will be executed. Rules in an active database system once are defined centrally and then will be used automatically during application execution. ADBS are useful for the application (e.g. warehousing, factories automation and complex financial calculations systems) that needs automatic monitoring.

One of the most important aspects of the ADBS that affects their power is rule language [15]. Besides the main components of rule language, there are additional features that play an important role in the specification of ECA rules. One of these important features is coupling modes. The phases of rule execution discussed so far are not necessarily executed contiguously, but depend on the so-called coupling modes which are pairs of values (x, y) associated with each rule. The value 'x' couples event signaling and condition evaluation of a rule, whereas 'y' couples condition evaluation and action execution [9]. Possible coupling modes are immediate, deferred and independent [15]:

- Immediate Mode: in this mode, when an event occurs, current transaction is suspended and the action is executed, if the condition holds.
- Deferred Mode: in this mode, after the occurrence of an event condition evaluation and action execution is deferred till the end of the current transaction. In deferred mode, the action of triggered rule should be executed before current transaction commits.
- Independent Mode: when an event is triggered in independent mode, there are no time-constraints and restrictions on condition evaluation and action execution.

In order to support reactive behavior of the ADBS, they should contain additional units for managing rule base and rule processing steps in comparing with passive database Management systems [10].

B. Rule Processing in Active Database System

In this part, we briefly describe what happens when an event occurs in the ADBS. An application is running sequentially until an event occurs. After an event occurs, the rule processing unit is activated and triggers the appropriate rule(s). Triggered rule(s) are queued in a temporary buffer. Then triggered rule(s) are selected according to some special criteria and then their "condition" section is evaluated. If condition is true, the action section will be executed. If the current rule triggers some other rules, new triggered rules will be passed to the rule processing unit. When there aren't any triggered rules, the application continues running [2, 9]. In summary, there are five different rule processing steps [9]:

- *Event Signaling*: When a primitive event occurs, the primitive event detector signals that event. Additionally, the composite event detector considers these primitive events that contribute to composite events.

- *Rule Triggering*: After the event is signaled, ECA rules that correspond to the signaled event are selected, and for each of them rule instances are created. In each rule instance, there is some additional information based on scheduling method, such as timestamp, deadline, execution time, etc. These rule instances are buffered to use in the next step.
- *Condition Evaluation*: After the buffering of rule instances, their conditions are evaluated. Then, for each rule with a true condition, a transaction is generated.
- *Transaction Selection*: This step is also called transaction scheduling phase. In this phase, a selection algorithm [16] operates on execution buffer and selects one transaction which is generated based on triggered rules, and sends the transaction to the execution unit.
- *Transaction Execution*: Transactions generated based on triggered rules are executed in this phase.

An example of database commands related to a stock exchange system is:

```
DEFINE LowRisk
ON Stock.UpdatePrice
IF (Stock.policy=Low_risk) and Stock.price> Stock.initprice )
DO Stock.Buy
```

In above example, a rule namely Low Risk has been defined and stock.Update.Price is an event which has occurred and the two conditions means that Stock.policy=Low_risk and Stock.price>Stock.initprice will be true, purchase the stocks is done automatically.

III. RULE SCHEDULING METHODS

For selecting one of the buffered rules the execution unit uses a selection algorithm. In this section we briefly describe the rules scheduling methods. Figure 1 shows the formal specification of a scheduling method in general [9].

- 1) $RULE_BASE \equiv \{Set\ of\ ECA\ RULE\}$
 - 2) $ACTIVE_RULE_BASE \equiv \{Set\ of\ ECA\ RULE\}$
 - 3) $Input \equiv RULE_BASE$
 - 4) $Output \equiv RULE$
 - 5) $n \equiv \{i \in N \mid i = \text{Number of ECA RULES in the } RULE_BASE\}$
 - 6) $ACTIVATE(R) \equiv \{Create\ Instance\ of\ R\}$
 - 7) $ADD_ACTIVE_RB(R) \equiv \{i \in N\ and\ i = n \mid ACTIVE_RULE_BASE[i+1] = R\}$
 - 8) $RULE_SELECTION(ACTIVE_RULE_BASE)$

Figure 1. The formal specification of scheduling method in general

In the following we will introduce and evaluate existing rule scheduling methods.

1) *Random Scheduling Method:* Random scheduling method is one of the simple methods for rules scheduling [14]. In this method, if in rules processing cycle we need a new rule to executing, then among the active rules, one rule will be randomly selected. The main advantage of this method is simple implementation but performance of random method is very low in comparison with other methods. Random scheduling method has been implemented in Ode and RPL active database systems [14].

2) *FCFS Scheduling Method:* In FCFS (First Come First Service) method, a timestamp that specifies when the rule is activated will be stored. At time of rule selection, the rule with minimum timestamp will be selected. This method is called timestamp method too. FCFS method increases the system performance, but this factor can't alone increase a great effect on ADBS performance. This method is used in SAMOS system [6].

3) *Static Priority Scheduling Method:* In this method the system assigns a numeric priority to each ECA rule that this priority should be unique. Then when an activated rule should be selected to run, the rule that has the minimum static priority is selected. For example, in the Ariel [12] and POSTGRES [13] systems is considered a priority between [-1000 to +1000] for each rule. Another version of static priority method is used of priority groups. In this method, each system have N priority group that regard to system structure, each of rules are put into the priority group.

4) *Parallel Scheduling Method:* In this method we can observe run of several rules simultaneously using properties of parallel execution. There are several buffers for new activated rules. Active rules with regard to timestamp of their activation are put into the buffers and will be run at suitable time. HipAc active database system supports parallel scheduling and executes all triggered rules concurrently [16]. There is another scheduling method based on parallel execution that using multithreading technique. The rules assigned to the threads and will be run parallel. This method has been implemented in FAR ADBS [4]. The main disadvantage of parallel scheduling method is that, it will not be used in all systems.

5) *EDF Scheduling Method:* EDF (Earliest Deadline First) method is one of the best methods of rules scheduling. This method is based on earliest dead line first and introduced for real-time active database systems [3] and the simulations show this method high performance in these systems. In this method, when an activated rule should be selected to execute, the rule that has the earliest dead line will be selected. This method has three different versions: (1) ED_{FPD}, (2) ED_{F_{DIV}}, and (3) ED_{F_{SL}}. The ED_{FPD} is a static baseline policy where rules priorities do not change with time. ED_{F_{DIV}} and ED_{F_{SL}} are dynamic policies where rules priorities change depending on the amount of dynamic work they have generated [3].

6) *Ex-SJF Scheduling Method:* In this method, when an activated rule should be selected to run, the rule that has the minimum execution time will be selected [9]. This method has three different versions: (1) EX-SJF_{EXA}, (2) EX-SJF_{PRO} and (3) EX-SJF_{PRO}. Practically, calculation real execution time of rules in run time is possible but due to high computational overhead, will be lead to an inefficient scheduling method. As a result, all versions of EX-SJF will calculate rules execution time before the rules are executed [11]. So for computing rule execution time before runtime, we should estimate the probability of the other rules in the rule body that activated at runtime dynamically (i.e. off springs of the rule). Execution of a rule depends on correctness of the condition part. In other word, execution probability of a rule is considered as condition correctness of that rule. Check the correctness of a conditional statement, before its execution will be difficult. Thus, all three versions of the EX-SJF method creates rules execution tree for predicting off springs of the rules. The difference between these versions is how to estimate probability of the rules execution.

IV. COMPARISON OF RULE SCHEDULUNG METHODS

A. Active Database System Simulator

For implementation rules scheduling methods and evaluate their performance, we need an environment that can simulate the active database system behavior. With such system we can implement each rule scheduling method and evaluate its performance. For this reason, an environment which is named Active Database System Simulator (ADSS) has been designed and implemented at the intelligent systems laboratory [11]. Figure 2 shows the architecture of the ADSS [11]. The ADSS has three major modules that are:

- *Object Management Unit:* In active database system simulator, data is defined by set of objects. Every data object have some predefined procedures that are available only through the same procedures. In fact, these procedures are commands that constituting the transaction. Furthermore, the object management unit is responsible for concurrency control of transactions [6]. Also, this unit informs rules management when data objects produce events.
- *Rules Management Unit:* This unit is responsible for protection, maintenance, activation, rules condition control and finally creating transaction from the rules action part. If rules management unit receive a message from objects management unit or transaction management unit based on occurring an event, then rule database will be evaluated to be determined based on the mentioned event, the rule or rules are activated or not [6, 11]. In addition, rules management unit controlled condition part of activated rules and if condition of each activated rules possess correct value, then produce a transaction based on action part of the rule.

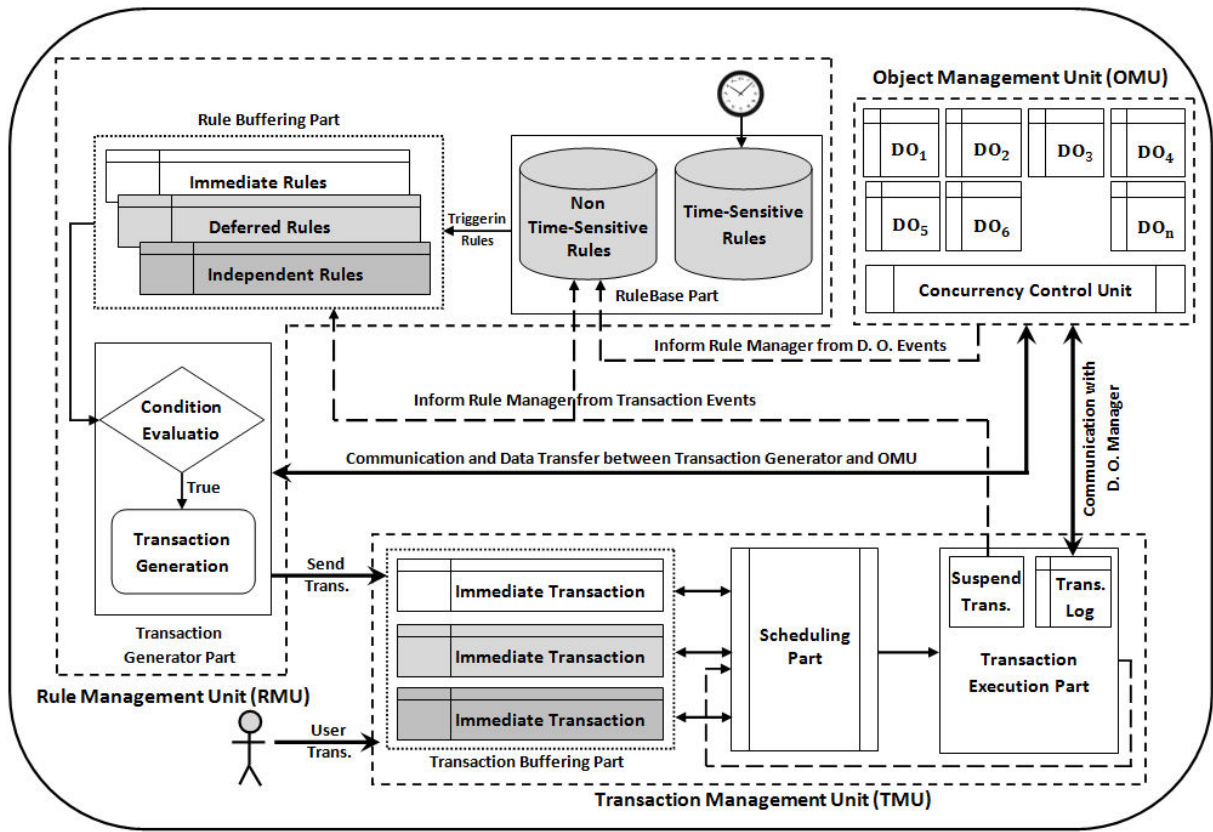


Figure 2. The active database system simulator

Rules management unit consist of three internal parts in the ADSS that are: Rulebase part, rule buffering part and transaction generator part.

- *Transaction Management Unit:* Transaction management unit is responsible for scheduling and execution of all transactions in the system. This unit includes three internal sections [11]. Transaction management unit consist of three internal parts in the ADSS that are: Transaction buffering part, scheduling part and transaction execution part.

B. Comparison of Scheduling Methods

In this section we introduce a framework for comparison and evaluation of existing rule scheduling methods. Figure 3 illustrates formal definition of evaluation criteria.

This framework contains five evaluation criteria: Average Response Time, Response Time Variance, Throughput, Time Overhead per Transaction and CPU Utilization [9]. Also as mentioned for implementation rules scheduling methods and evaluate their performance, we need an Active Database System Simulator (ADSS) that can simulate the active database system behavior. Experiments are performed in three modes [9]: (1) Deferred mode, (2) Immediate mode and (3) Composite mode. In the first mode system uses rules only in deferred mode. In the second mode, system uses rules only in immediate

mode and ultimately in the third mode, system uses rules in all immediate, deferred, and independent modes.

$$\begin{aligned}
 & \text{ART} = \text{Average Response Time} \\
 & \text{RTSV} = \text{Response Time Standard Variance} \\
 & \text{UCPU} = \text{CPU Utilization} \\
 & \text{TOPT} = \text{Time Overhead Per Transaction} \\
 & T_1^i = \text{Activation Time of } i^{\text{th}} \text{ Rule} \\
 & T_2^i = \text{Start of Execution Time of } i^{\text{th}} \text{ Rule} \\
 & T = (T_2^N + \text{Execution Time of } N^{\text{th}} \text{ Rule}) - T_1^1 \\
 & T^* = \sum_{i=1}^N \text{Real Execution Time of } i^{\text{th}} \text{ Rule} \\
 & U_{\text{CPU}} = \frac{T^*}{T} * 100, \quad \text{RSTV} = \sqrt{\frac{\sum_{i=1}^N ((T_2^i - T_1^i) - \text{ART})^2}{N}} \\
 & \text{ART} = \frac{\sum_{i=1}^N (T_2^i - T_1^i)}{N}, \quad \text{Throughput} = \frac{N}{T}, \quad \text{TOPT} = \frac{T - T^*}{N}
 \end{aligned}$$

Figure 3. Formal definition of evaluation criteria

Table 1 illustrates the experimental results. The Results shows that the Ex-SJF_{PRO}-V.1.8 has generally the most positive impact on performance (Response Time, Response Time Variance, and Throughput, Time Overhead per Transaction and CPU Utilization) of ADBS. In [9] evaluation results of different methods are discussed in detail.

TABLE I. EVALUATION RESULTS OF SCHEDULING METHODS

Evaluation Criteria Methods	Average Response Time	Response Time Variance	Throughput	Time Overhead Per Transaction	CPU Utilization
Random	3	1	5	4	3
FCFS	3	1	3	3	3
Static	3	1	4	4	3
Parallel	3	1	2	4	3
EDF _{PD}	3	1	2	6	3
EDF _{DIV}	3	1	2	3	3
EDF _{SL}	1	1	3	5	3
Ex-SJF _{EXA}	2	1	2	5	2
Ex-SJF _{PRO}	1	2	2	2	2
Ex-SJF _{PRO} -V. 1.8	1	2	1	1	1

CONCLUSION

Traditional database management systems (DBMS) have static nature. Means that the operations such as query, update and etc is performed only when asked by the user and DBMS doesn't act initiative in the specific condition. But many applications need to have automated monitoring, till in the occurrence specific event database management system will perform suitable reaction automatically. For this reason, active database systems (ADBS) were designed. In this paper, at first active database management system was introduced. Then, we investigated one of the most fundamental issues in active database system namely rules scheduling. And then different methods of rules scheduling has been studied. Finally, we compared different methods of rules scheduling using an Active Database System Simulator (ADSS).

REFERENCES

[1] E. Bertino, B. Catania and G. P. Zarri, "Intelligent Database Systems", Addison-Wesley, 2001.

[2] E. Hanson, U. Dayal and J. Widom. "Active Database Systems", Jennifer Widom, 1994.

[3] R. M. Sivasankaran, J. A. Stankovic, D. Towsley, B. Purimetla and K. Ramamritham, "Priority Assignment in Real-Time Active Databases", The International Journal on Very Large Data Bases, 5(1), 1996.

[4] S. Ceri, C. Gennaro, S. Paraboschi, G. Serazzi, "Effective Scheduling of Detached Rules in Active Databases", IEEE Transaction Knowledge and Data Engineering, 15(1), 2003.

[5] M. Stonebraker and G. Kemnitz, "The POSTGRES next generation database management system", Communications of the ACM, 34(10): 78-92, 1991.

[6] A. Geppert, S. Gatzju, K. R. Dittrich, H. Fritschi, and A. Vaduva, "Architecture and implementation of the active object-oriented database management system SAMOS", Technical Report 95.29, CS Department, University of Zurich, 1995.

[7] E. N. Hanson and J. Widom, "An Overview of Production Rules in Database Systems", In the Knowledge Engineering Review, 8(2):121-143, 1993.

[8] J. Widom, "The Starburst Rule System", in Active Database Systems: Triggers and Rules for Advanced Systems, Morgan-Kaufman Publishers, Sanfrancisco (Calif), 1996.

[9] A. Rasoolzadegan, R. Alesheykh and A. Abdollahzadeh, "Measuring Evaluation Parameters in Benchmarking Rule Scheduling Methods in Active Database Systems", IEEE International Conference on Computer and Communication Engineering (ICCCE'06), Kuala Lumpur, Malaysia, May 2006.

[10] A. Vadua, "Rule Development for active database", PhD Thesis, CS Department, University of Zurich, 1999.

[11] A. Rasoolzadegan and A. Abdollahzadeh, "ADSS: Active Database System Simulator for Adjustment Comparison of Scheduling Methods", 15th National Conference on Electrical Engineering, Iran Telecommunication Research Center, 2007.

[12] J. Stankovic, M. Spuri, M.D. Natale and G.C. Buttazo, "Implications of Classical Scheduling Results for Real-Time Systems", IEEE Computer Society Press, Los Alamitos, CA, 1995.

[13] S. Potaminsto and M. Stonebraker, "The POSTGRES Rule System", in Active Database Systems: Triggers and Rules for Advanced Systems, Morgann Kaufmann Publishers, Sanfrancisco, CA, 1996.

[14] S. Chakravarthy, "Architectures and monitoring techniques for active databases: An evaluation", In Technical Report TR-92-041, University of Florida, Gainesville, 1992.

[15] H. Fritshi and Z. Flaach, "A Component Framework to Construct Active Database Management Systems", PhD Thesis, CS Department, University of Zurich, 2002.

[16] Vaduva, S. Gatzju and K. R. Dittrich, "Investigating Termination in Active Database Systems with Expressive Rule Languages". In Proceedings of the 3rd International Workshop on Rules In Database Systems, pp. 149-164, Skovde(Sweden), 1997.

[17] C. J. Date, "An Introduction to Database Systems", 8th Edition, Addison-Wesley, 2003.

Analysis of Inbound and Outbound Email Traffic and its SPAM Impact

Seema Khanna

National Informatics Center
New Delhi, India
seema@gov.in

Harish Chaudhry

Department of Management Studies
Indian Institute of Technology
New Delhi, India
hciitd@gmail.com

Abstract— Spam is a huge issue for most Internet users – in fact, 52% of participants polled in a recent survey stated that spam was a major problem. And despite the evolution of anti spam software, such as spam filters and spam blockers, the negative effects of spam are still being felt by individuals and businesses alike. It is an examination of the various spam techniques and types of popular email viruses.

This work intends to increase the awareness and understanding of the spam emails that are received by internet user throughout the world. This paper analyse inbound and outbound email data from a particular organization to provide some hard figures and gain an in-depth understanding for the spamming and various virus types. A few spam control measures at the users' level are also suggested. This conceptual paper is definitely expected to contribute to future research on similar and related topics as spin off from this study.

Keywords-email; traffic; inbound; outbound; spam;

I. INTRODUCTION

Electronic mail, commonly known as email or e-mail, is a method of exchanging digital messages from an author to one or more recipients. Modern email operates across the Internet or other computer networks. Some early email systems required that the author and the recipient both be online at the same time, in common with instant messaging. Today's email systems are based on a store-and-forward model. Email servers accept, forward, deliver and store messages.[1]

Email is usually considered to be a one-to-one communication medium, but at the Internet Service Provider (ISP) level, many email flows are mailing-lists (one-to-many) or forwarded traffic (many-to-one). However, the literature contains few, if any, quantitative measures of what is meant by many". [2]

II. OBJECTIVES

- The overriding objective is to analyse inbound and outbound email data from a particular organization to provide some hard figures.
- The secondary objective is to and gain an in-depth understanding for the spamming and exhaustively explore the various types of email viruses.

III. STATISTICS

Spam accounts for 14.5 billion messages globally per day. In other words, spam makes up 45% of all emails. Some research companies estimate that spam email makes up an even greater portion of global emails, some 73% in fact. The United States is the number one generator of spam email, with Korea clocking in as the second largest contributor of unwanted email.

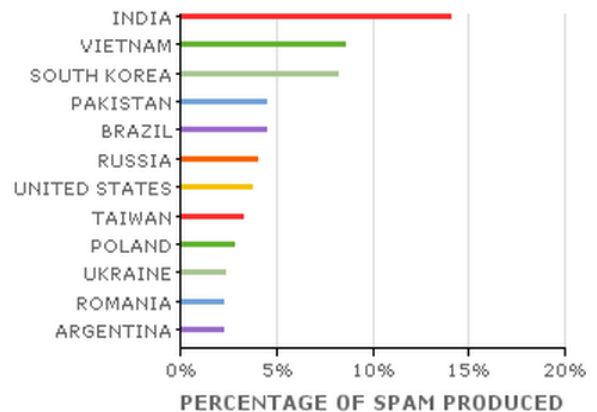


Figure 1: SPAM Sources by country[3]

The most prevalent type of spam is advertising-related email; this type of spam accounts for approximately 36% of all spam messages. The second most common category of spam is adult-related in subject and makes up roughly 31.7% of all spam. Unwanted emails related to financial matters is the third most popular form of spam, at 26.5%.

Surprisingly, scams and fraud comprise only 2.5% of all spam email; however, identity theft (which is known as phishing) makes up 73% of this figure. Because spam has inundated both the personal and corporate world of emailing, it has affected the way that individuals and companies feel about spam. In fact, surveys have found that spam has led to decreased public confidence and trust in Internet communications.

Companies also find spam a troublesome problem that reduces productivity and safety. Fifty two percent of companies interviewed for a recent study listed minimizing spam as their top priority. However, anti spam measures such as spam

blockers provide some hope in the fight against unwanted email. In fact, MSN alone blocks some 2.4 billion spam emails every day.

According to a study by the Radicati Research Group Inc., a research firm based in Palo Alto, California, spam costs businesses \$20.5 billion annually in decreased productivity as well as in technical expenses. Nucleus Research estimates that the average loss per employee annually because of spam is approximately \$1934.[4]

Predictions for the future costs of spam don't look any brighter. It is estimated that 58 billion junk emails will be sent every day within the next four years, a figure that will cost businesses some \$198 billion annually. However, some researchers believe that based on an estimated current cost of \$49 annually per inbox, the total cost of spam for businesses will balloon to \$257 billion per year if spam continues to flourish at its current rate.

IV. TRAFFIC ANALYSIS

Traffic analysis is the process of intercepting and examining messages in order to deduce information from patterns in communication. It can be performed even when the messages are encrypted and cannot be decrypted. In general, the greater the number of messages observed, or even intercepted and stored, the more can be inferred from the traffic. Traffic analysis can be performed in the context of military intelligence or counter-intelligence, and is a concern in computer security. Traffic analysis is also a concern in computer security. An attacker can gain important information by monitoring the frequency and timing of network packets. A timing attack on the SSH protocol can use timing information to deduce information about passwords since, during interactive session, SSH transmits each keystroke as a message. [5] The time between keystroke messages can be studied using hidden Markov models. Song, et al. claim that it can recover the password fifty times faster than a brute force attack.

A. INBOUND Traffic

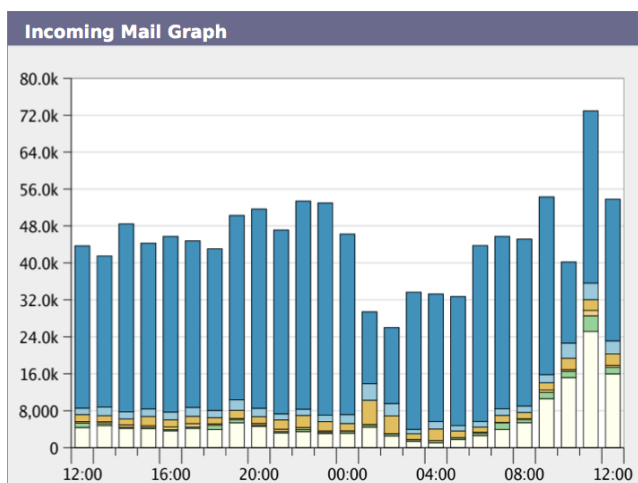


Figure 2. Incoming Mail Graph

Message Category	%	Messages
Stopped by Reputation Filtering	76.4%	858,874
Stopped as Invalid Recipients	4.2%	47,174
Spam Detected	4.4%	49,050
Virus Detected	0.0%	90
Stopped by Content Filter	0.8%	8,549
Total Threat Messages:	85.8%	963,737
Marketing Messages	1.6%	17,893
Clean Messages	12.7%	142,228
Total Attempted Messages:		1,123,858

B. OUTBOUND Traffic

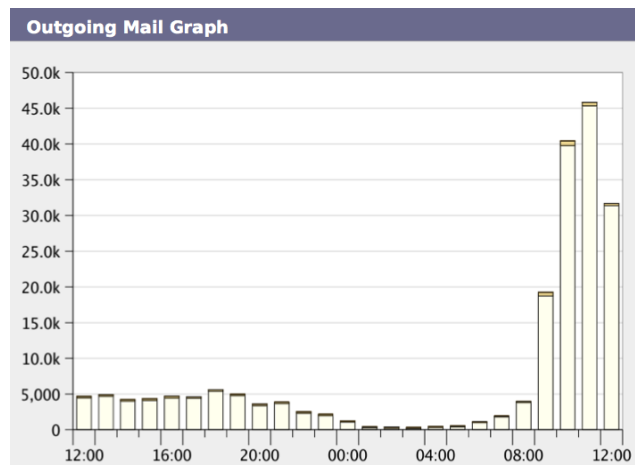


Figure 3. Outgoing Mail Graph

Message Processing	%	Messages
Spam Detected	0.1%	114
Virus Detected	0.0%	9
Stopped by Content Filter	3.0%	5,893
Clean Messages	97.0%	191,807
Total Messages Processed:		197,823

V. SPAM, SPAM, SPAM

Spam is the use of electronic messaging systems to send unsolicited bulk messages indiscriminately. While the most widely recognized form of spam is e-mail spam, the term is applied to similar abuses in other media: instant messaging spam, Usenet newsgroup spam, Web search engine spam, spam in blogs, wiki spam, online classified ads spam, mobile phone messaging spam, Internet forum spam, junk fax transmissions, social networking spam, television advertising and file sharing network spam.[6]

Spamming remains economically viable because advertisers have no operating costs beyond the management of their mailing lists, and it is difficult to hold senders accountable for their mass mailings. Because the barrier to entry is so low, spammers are numerous, and the volume of unsolicited mail has become very high. In the year 2011, the estimated figure for spam messages is around seven trillion. The costs, such as lost productivity and fraud, are borne by the public and by

Internet service providers, which have been forced to add extra capacity to cope with the deluge. Spamming has been the subject of legislation in many jurisdictions[7]. A person who creates electronic spam is called a spammer[8].

A. How do spammers get your email address ?

The Internet is full of resources which can be used, directly or indirectly, as databases of email addresses. These include publicly available mailing list archives, USENET, web pages, member directories and search engines. Spammers obtain email addresses, using various techniques including manual and automated software also known as crawlers, which crawl the web searching out email addresses.

1) Search Engines

Queries on Internet search engines can be used to search email addresses. Spammers use manual and automated crawlers to harvest email addresses from search engines. Software that can extract and collect email addresses from different search engines is also available.

2) Directory Harvest Attack/ Brute Force Attacks

A Directory Harvest Attack or DHA is a technique used by spammers in an attempt to find valid/existent e-mail addresses at a domain or by using brute force. The attack is usually carried out by way of a standard dictionary attack, where valid e-mail addresses are found by brute force guessing valid e-mail addresses at a domain using different permutations of common usernames. In most of the email deployments today the SMTP gateway appliances are integrated with a LDAP. The recipient in a mail is checked against the LDAP and if valid, mail is accepted.[9]

3) Address posted in public domain/forum

Email addresses posted on websites are collected from the HTML pages by spammers using manual and automated crawlers. Spammers use programs which spider through web pages, looking for email addresses, e.g. email addresses contained in MAILTO: HTML tags. Email addresses published on the Internet usually receive a huge amount of spam.

4) Chat Rooms

Chat rooms are another major source of email addresses for spammers. Email addresses can be easily obtained in some IRC channels and other chat rooms. Spammers harvest these email addresses, knowing that these are 'live' addresses. IRC bots are also used to send messages interactively to IRC and to chat rooms to harvest email addresses.

5) Readymade Lists

Readymade lists of email addresses are available for purchase on the Internet or otherwise. Spammers are also said to sell or exchange readymade email address lists collected through various means.

6) Domain Name Registration

Usually, every domain has three contact points - administration, technical, and billing. The contact points include the email address of the contact person. The list of domains are usually made available to the public by the domain registries.

B. SPAM Techniques

1) SMTP Loopholes

The claimed source identity in a spam message can either be authentic or false. Some users or organizations send bulk mail with their original identity, without forging the sender name and domain name.

However, in most cases, the sender of the spam forges the user name and domain name. Spammers also try to impersonate as legitimate users or organizations. The forging of identity or impersonation of identity in electronic mailing systems is done by exploiting a weakness in the SMTP protocol and its implementation. SMTP does not include any sender authentication to verify the authenticity of the sender.

2) Open Relay

Internet was designed for redundancy. In its original design, an SMTP server was designed to accept mail destined for any other domain and forward it appropriately. This facility is known as relaying. However, this facility began to be misused by spammers.

When there is no authentication mechanism to prevent such misuse of the relay and any sender can send his mail to another domain using the Mail relay, it is called an 'Open relay'. An Open Relay doesn't restrict any client from forwarding mails to another domain through it.

3) Stealth and Open Proxies

A Proxy Server should accept requests only from its own clients by either forcing a client to connect only from a range of IP addresses, or by using authentication. Any proxy server that doesn't restrict its client base to its own set of clients and allows any other client to use it is known as an Open Proxy.

Open Proxies are used by spammers for mass mailing using forged identities. An Open Proxy can be used by a spammer to anonymously connect to a mail server. Further, any mail that a spammer sends shall appear to originate from the Open Proxy server. The use of an Open Proxy can also be used to bypass filters based on both domain name and IP address.

4) Mass mailing worms

Mass-mailing worms are a general class of malicious code which propagate through email. They generate high volumes of traffic that clog networks and overload mail relays. A lot of spam originates from home computers infected with various mass mailing worms. According to one estimate, about Four-fifths of spam is a result of these mass mailing worms.

5) 'Spam Zombie' Machines

Spammers exploit these systems by turning them in to 'Spam zombie' machines, programmed to send out spam. According to some estimates, about one-third of spam originates from these zombie computers.

VI. VIRUS TYPES

A computer virus is a computer program that can replicate itself[10] and spread from one computer to another. The term "virus" is also commonly, but erroneously, used to refer to other types of malware, including but not limited to adware and spyware programs that do not have a reproductive ability.

Threat Category	Messages
Phish	11
Scam	11
Virus	22738
Total Messages	22760

Table 1. Threat Types

A virus comes within an attached file in an e-mail message. When that file is opened, the virus does its damage. Macro viruses can come in Microsoft Word documents that are sent as e-mail attachments. The macro causes the damage when the document is opened providing macro processing has not been disabled within the Microsoft Word application.[11].

Figure 4 show various incoming virus types detected on an Email Security Appliance.

Virus Type	Incoming	Outgoing	Total Infected Messages
W32/Mydoom.o@MM!zip	55	0	55
W32/Mydoom.o@MM	10	0	10
XF/Sic.gen	0	8	8
Mal/BredoZp-B	4	0	4
Troj/Agent-WER	4	0	4
W32/MyDoom-O	4	0	4
W32/Netsky.p@MM	3	0	3
Troj/ZipMal-AW	2	0	2
Mal/KeyGen-M	1	0	1
Mal/Phish-A	1	0	1

Table 2. Virus Type Detail

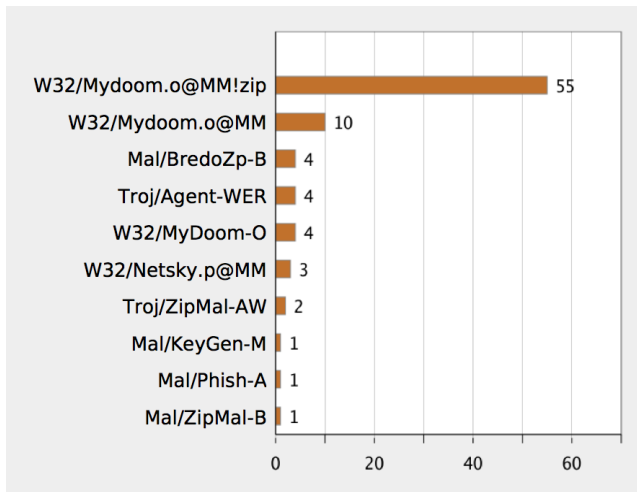


Figure 4. Incoming Virus Type

Figure 5 show various outgoing virus types detected on an Email Security Appliance.

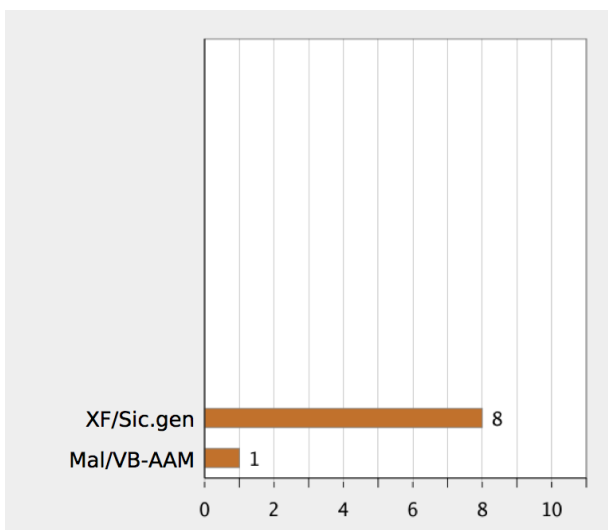


Figure 5. Outgoing Virus Type

VII. SPAM FILTERING

Email spam, also known as junk email or unsolicited bulk email (UBE), is a subset of electronic spam involving nearly identical messages sent to numerous recipients by email. Definitions of spam usually include the aspects that email is unsolicited and sent in bulk.[12][13][14].

Spammers collect email addresses from chat rooms, websites, customer lists, newsgroups, and viruses which harvest users' address books, and are sold to other spammers. Much of spam is sent to invalid email addresses. Spam averages 78% of all email sent.[15] According to the Message Anti-Abuse Working Group, the amount of spam email was between 88–92% of email messages sent in the first half of 2010.[16]

A. Content-based filtering

Content based filtering does a detailed inspection on the contents of an email message and helps in identifying spam messages. Content based filtering can be implemented through simple text pattern matching, or through statistical probability indication.[17]

1) Pattern matching

Content based filtering depends on predefined patterns of text and rule-based rankings. They evaluate a large number of patterns against a candidate message. Some matched patterns add to a message score, while others subtract from it. The incoming e-mail is evaluated based on simple strings found in specific header fields, the header in general, and in the email body itself. If the score of a message exceeds a certain threshold, it is filtered as spam, otherwise it is considered a legitimate mail.

2) Hash matching

A database of hashes of known spam messages is stored. Each new email received is hashed and compared with the above database. If the hash matches any of the stored hash values, it is identified as spam.

3) Statistical Classification Engines

Spam filtering can also be performed using statistical classification of the contents of the message. It is said to be one

of the most effective spam fighting methods. These engines assign a spam indicative mark to words or chunk of words based on previously identified spam messages. New incoming messages are verified against these to generate spam indicative probability. Thus, based on the prior appearance of certain words, or chunk of words, statistical classification engines determine the probability that the new email message is spam. Statistical engines build spam indicative probabilities of words automatically, with minimal human intervention. One of the most popular statistical classification engines is Bayesian filter.

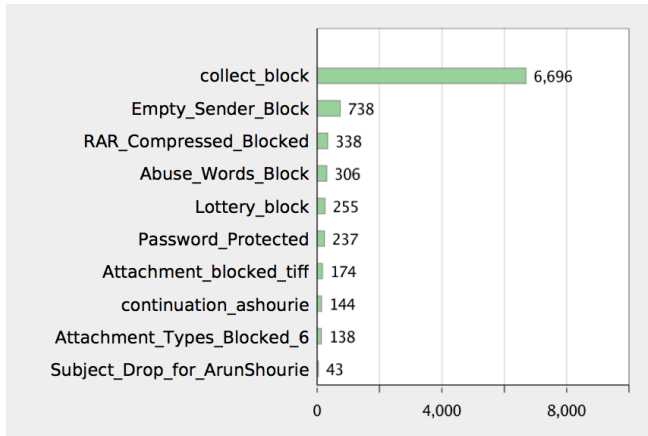


Figure 6. Top Incoming Content Filter Matches

Table 3. shows the Incoming Content Filter Matches analyzed over a period of 4weeks in a particular organization.

Content Filter	Messages
collect_block	6696
Empty_Sender_Block	738
RAR_Compressed_Blocked	338
Abuse_Words_Block	306
Lottery_block	255
Password_Protected	237
Attachment_blocked_tiff	174
continuation_ashourie	144
Attachment_Types_Blocked_6	138
Subject_Drop_for_ArunShourie	43
Specific_Subject_Blocked	31
Blocked_Senders_for_Arun	17
Attachment_blocked_exe	15
IP_Address_in_ID_Blocked	2
citicorp_statement	2
Blocked_Senders_for_Jaswant	1
Total Incoming Matches	9137

Table 3. Incoming Content Filter Matches

4) Heuristic Filters

Some filtering mechanisms implement a combination of the above techniques i.e. pattern matching and Bayesian Filters. Heuristics are a series of rules used to score the spam probability of an email. These are human-engineered rules by

which a program analyses an email message for spam-like characteristics. These rules may look for multiple uses of certain phrases incorporating hundreds of rules in order to catch spam. A message might get a certain number of points for containing a certain phrase; more points if it contains a URL link, and even more points if the message includes a phrase for un-subscription request link. Depending on the parameters established, reaching a certain score would classify the message as spam

Heuristic engines are quite effective though the spammers are using more sophisticated techniques to get around this kind of filtering. Spammers reverse-engineer heuristic rules and create messages that can bypass the filters.

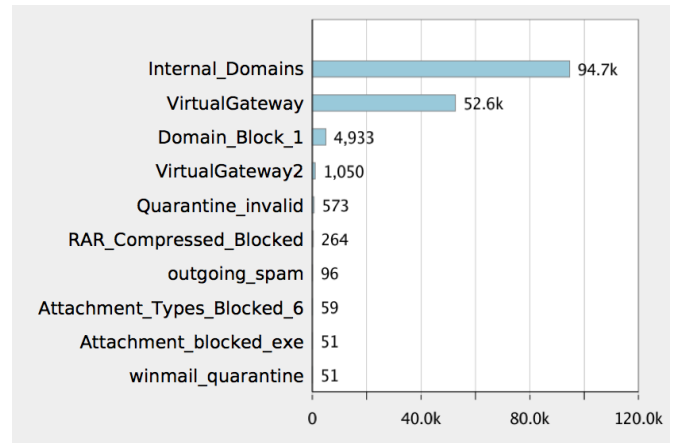


Figure 7. Top Outgoing Content Filter Matches

Table 4. shows the Outgoing Content Filter Matches analyzed over a period of 4weeks in a particular organization.

Content Filter	Messages
Internal_Domains	94682
VirtualGateway	52644
Domain_Block_1	4933
VirtualGateway2	1050
Quarantine_invalid	573
RAR_Compressed_Blocked	264
outgoing_spam	96
Attachment_Types_Blocked_6	59
Attachment_blocked_exe	51
winmail_quarantine	51
Bypass_Quarantine_Invalid	47
Abuse_Words_Block	38
Lottery_Block	34
Attachment_blocked_tiff	22
Barc	10
Block_Attachements	8
Total Outgoing Matches	154562

Table 4. Outgoing Content Filter Matches

B. Source address-based filtering

Source-based filtering is overtaking content-based filtering as the main method for blocking spam. Spammers are always figuring out new ways to get around content filters, but hiding the source IP address and its behavior, is more difficult.

1) White list/verification filters

A white-list contains a list of source addresses, which the recipient wants to receive mail from and are approved by the receiver as not being sources of spam. A white-list filter works based on the white-list of explicitly approved source addresses. Mails originating from sources, defined in the white list, are forwarded on to the mailbox and the rest tagged as spam. The disadvantage is that they place an extra burden on legitimate senders. An effort is also required to receive mails from new users.

2) Blacklists

A blacklist is almost opposite of a white list. It contains a list of source addresses which are known to be originators of spam. Local blacklists are usually prepared and maintained by administrators manually. However, global blacklists of well known spammer IPs, and domain names are maintained by different organizations, and are termed as distributed black lists.

Reputation analysis of mail sending IP addresses is another method to filter spam. Amount of spam received from an IP address is monitored and after a certain threshold it is recognized as a spam source and future attempts of that IP address to send mail will be blocked. This is a method to maintain blacklists.

3) Realtime BlackHole Lists (RBLs)

Realtime BlackHole Lists (RBLs) can be queried by SMTP engines to verify any incoming messages. RBLs contain IP addresses which are known to be originator of spam. RBLs are normally implemented by using a protocol similar to DNS and are popularly known as DNSRBLs. RBLs are maintained by many different RBL operators, and organizations can simply subscribe to them.

Spam is also identified through Mail Header analysis. An e-mail message contains routing information which can be analyzed to determine discrepancies in the format, because many spammers try to hide their tracks by placing invalid information in the header. The discrepancies may indicate a possible spam.

Possibility of false positive (i.e. blocking legitimate emails) is a major concern in spam filtering. The disadvantage of source address based filtering is that they can be cumbersome and time-consuming; requiring constant list maintenance in order to be effective. Spammers also use hundreds of new addresses, acquire new zombies or relay machines to route their spam. Sometimes spammers spoof legitimate senders addresses adding more difficulties to address-based filtering.

C. Challenge-Response

Like white lists, challenge-response systems allow mail only from previously authenticated address lists. Mails from new addresses are kept on hold and a challenge mail is sent

back to the sender. If the sender successfully replies to the challenge, the mail is accepted as authentic mail and delivered to the receiver's inbox. The challenge message may contain some simple questions, which can only be replied by human users.

This mechanism nullifies the possibility of spam being sent from an automated spamming tool like different worms, zombie machines etc. Though it is an effective method to block massive spam being generated by automated tools, worms etc, it puts significant overhead to the legitimate mail senders. Mails from automated systems like webpage re-mailers cannot be delivered to such systems. Even legitimate emails may not be delivered because the challenge is not completed. Deadlock situation may also arise when both the recipient and sender use challenge response systems to authenticate each other.

VIII. SPAM CONTROL

A. Best Practices

1) SMTP Server Implementation

- Should not relay unauthorized mails (Should not be an Open Relay)
- Separate ports for submission and relay of messages
- Implement client authentication before mail submission
 - POP before SMTP
 - SMTP-Auth with TLS (RFC 2554)
- Disable SMTP commands like VERIFY (VRFY)
- Prevent remote mails to local groups
- Define maximum number of recipients per message
- Reject NULL sender identity
- Maintain statistical analysis
 - Top SMTP connections from hosts/IPs
 - Top "From Addresses"
 - Top "Recipient Address"

2) Internet Service Providers

- Close All Open Relays
- Control Open Proxy Servers
- Control hosts which can act as an SMTP server for mail transaction
- Control home SMTP traffic
- Implement Rate Limits on Outbound Email Traffic
- Monitor and control web based E-Mail submission forms
- Detect and contain virus/worm infected machines

- Logging of sources of mails generated within their network
- Disseminate information and Educate Customers

3) *Email Users*

- Disguise e-mail addresses posted in a public electronic place.
- Carefully subscribe to commercial e-mail news letter
- Use multiple e-mail addresses
- Disposable email address
- Use a filter
- Short e-mail addresses

4) *Code of Ethics for mass mailers/advertisers*

Advertising via the use of Mass-Messaging ‘Spam’ is financially very attractive, as very little cost is used in sending spam. However, as a responsible Internet user, advertisers should avoid this option considering the various risks associated with it. The risks associated with mass mailing by advertisers range from bad publicity to legal ramifications to ISP service discontinuity. RFC 3098: How to Advertise Responsibly Using E-Mail and Newsgroups provides a framework through which an advertiser, the recipients and the Internet community can coexist in a productive and mutually respectful fashion.

B. *Long Term Measures*

Various technical controls discussed to minimize spam are not sufficient to control it completely. Despite all the controls, spam is growing every year. Spammers are continuously innovating to beat the controls. The Internet can not be completely spam free unless and until the basic flaws of SMTP protocol are eradicated. Restructuring of SMTP protocol to address the flaws is considered to be the long term measure to control spam. Various frameworks and methodologies are being proposed by different research groups to address the SMTP design flaws.

1) *Sender Policy Framework*

Sender Policy Framework is an extension of SMTP designed by Meng Weng Wong of POBOX. The framework suggested a mechanism for E-Mail source Identity validation. [18]. SPF was submitted as input along with other proposals to the IETF MARID workgroup. The latest version of SPF is known as the Marid Proposal.

SPF allows an Internet domain to specify which machines are authorized to send e-mail for that domain. This information is specified using ‘reverse MX’ entries in Domain Name Service (DNS) records. The receiving mail servers implementing SPF then treat as spam any email that claims to originate from a domain but fail to originate from the servers authorized to send mail for that domain.

The SPF operates at the level of the SMTP transaction. It verifies i) The MAIL FROM: parameter of the incoming mail, ii) the HELO/EHLO parameter of the sending SMTP server and iii) the IP address of the sending SMTP server.

SPF has some limitations. It only validates the domain of the envelope sender (listed as "Return-Path: " in e-mail headers). Thus, domains that share mail senders could forge each others' domain. SPF does not validate that a given email actually came from the claimed user. SPF also breaks the inter-system SMTP forwarding (where an agent forwards email to someone else without changing the "from" address).

2) *Sender ID*

Sender ID is a proposed specification developed within the MARID IETF Working Group between May and October 2004. The Sender ID framework is an extension of previous SPF proposal. It consists of two halves: the SPF Classic half, and the Purported Responsible Address (PRA) half.[19]

This combined specification is the result of Microsoft's Caller ID for E-Mail proposal, Meng Wong's Sender Policy Framework (SPF), and a third specification called the Submitter Optimization.

Purported Responsible Address (PRA) is determined by an algorithm which determines the source of the email based on the email headers. While SPF is intended to work in the MTA level, PRA is checked at MUA levels. It is possible to authenticate this email by tracking the PRA.

The PRA is as correct as the email headers. "Purported" reveals the claimed source of a message from the headers and not necessarily where it actually came from.

3) *Domain Keys*

Domain Keys framework is an Internet-Draft for publication to the Internet Engineering Task Force (IETF). Domain Keys uses the concept of digital signatures. There are two steps to signing an email with Domain Keys:

- On the sender side, the domain owner generates a public/private key pair to use for signing all outgoing messages. The public key is published in the DNS, and the private key is made available to their Domain Key-enabled outbound email servers. When each email is sent by an authorized end-user within the domain, the Domain Key-enabled email system automatically uses the stored private key to generate a digital signature of the message. This signature is then pre-pended as a header to the email, and the email is sent on to the target recipient's mail server.
- On the receiving side, the receiving email system extracts the signature and fetches the public key from DNS for the claimed From: domain. The Public key is used to verify the signature. If the domain is verified the email is delivered to the user's inbox.

4) *Cisco's Identified Internet Mail*

Identified Internet Mail (IIM) has been proposed by Cisco Systems as a signature-based mail authentication standard designed to address spam. IIM allows a recipient to authenticate messages and verify their authorization. IIM verifies that the message sender is authorized to send messages using a given e-mail address and that the original message was not altered in any consequential manner. Identified Internet Mail messages are in standard RFC 2822 format with

additional headers called IIM-Signature and IIM-Verification. An Internet-Draft of IIM (draft-fenton-identified-mail-00) was submitted to the IETF. Cisco Systems has released an open source implementation of Identified Internet Mail [20].

C. Legislative Measures

Technical controls alone can never eradicate spam completely. There is a need to define spam as a crime and bring it under the purview of legislation. The United States and various other countries have already adopted anti-Spam laws. Under these laws, commercial unsolicited mails are prohibited and rules have been framed for sending commercial e-mails.

A spam law should enforce the following:

- UCE messages must not be sent
- Commercial electronic messages must include accurate sender information
- Commercial electronic messages must contain a functional unsubscribe facility
- Address-harvesting software & harvested lists must not be supplied, acquired or used.
- Penalties & compensation for spamming

India, as on date doesn't have an Anti-Spam Act (loosely; Information Technology Act, 2000), though some prosecution has already taken place using existing legislation. One of the most successful spam laws to date is said to be the Australian spam Act that went into effect in the year 2004. Australian spam Act enforces three key elements - Consent, Identify and Unsubscribe. Australian spam Act enforces the concept of opt-in lists. Nearly a year after the passage of the US Federal Can-Spam Act, the law has not been able to curb spam, according to a report published by MX Logic, an anti-Spam company.[21]

CONCLUSION

It is a cat-and-mouse game. Due to the social and human components, there are no completely effective solutions. Only through learning from our shared experiences can we hope to better protect Internet users. Spammers are getting better every day. The anti-spam industry has taken up the challenge and today we have multiple solutions to the problem. There will be new advancements and changes in the techniques that are being used by the attackers, these developments are to be carefully analyzed and are to be dealt in a very thoughtful manner. We need to move towards effective solutions with proper legislative laws and at the same time without overburdening the user.

None of the spam control discussed in this paper is sufficient to make Internet spam free. Industry wide collaboration, cooperation and knowledge sharing is required towards this. Initiative to adhere to the best practices by all service providers and end users is also required. We suggest limiting some of the features allowed and following best practices as well as spam controlling mechanisms for better protection when working in critical environments where security is a priority.

ACKNOWLEDGMENT

The work of any individual is a reflection of the support, influence, and inputs received from a multitude of others. Recognition is due to many people for their suggestions and encouragement; to attempt to name each would run the very real risk of excluding one or more. My sincere appreciation is due to all those in my life who have provided encouragement and have steadily supported every endeavor with advice and assistance. A big thank you to all my seniors for their specific advice and counsel. It would not be out of place to thank all the authors and researchers whose work I have consulted for writing this paper.

REFERENCES

- [1] Wikipedia, "Email". <http://en.wikipedia.org/wiki/Email>. Accessed: December 21, 2011.
- [2] Richard Clayton, "Email Traffic: A Quantitative Snapshot"
- [3] M86 Security Labs, "Spam Statistics: Spam Sources by Country". Statistics for Week ending May 20, 2012. http://www.m86security.com/labs/spam_statistics.asp. Accessed: May 25, 2012.
- [4] Spam Laws, "Spam Statistics and Facts". <http://www.spamlaws.com/spam-stats.html>. Accessed: December 21, 2011.
- [5] Song, Dawn Xiaodong; Wagner, David; Tian, Xuqing (2001). Timing Analysis of Keystrokes and Timing Attacks on SSH. 10th USENIX Security Symposium
- [6] Wikipedia, "Spam (Electronic)". [http://en.wikipedia.org/wiki/Spam_\(electronic\)](http://en.wikipedia.org/wiki/Spam_(electronic)) Accessed: December 21, 2011.
- [7] The Spamhaus Project - The Definition Of Spam
- [8] Gyongyi, Zoltan; Garcia-Molina, Hector (2005). "Web spam taxonomy". Proceedings of the First International Workshop on Adversarial Information Retrieval on the Web (AIRWeb), 2005 in The 14th International World Wide Web Conference (WWW 2005) May 10, (Tue)-14 (Sat), 2005, Nippon Convention Center (Makuhari Messe), Chiba, Japan.. New York, N.Y.: ACM Press. ISBN 1-59593-046-9
- [9] Seema Khanna, Dr. Harish Chaudhry "Anatomy of compromising email accounts", IEEE International Conference on Information and Automation (ICIA12), China. To be published: June 2012.
- [10] Dr. Solomon's Virus Encyclopedia, 1995, ISBN 1-897661-00-2, Abstract at <http://vx.netlux.org/lib/aas10.html>
- [11] <http://encyclopedia2.thefreedictionary.com/e-mail+virus>
- [12] James John Farmer (2003-12-27). "3.4 Specific Types of Spam" (FAQ). An FAQ for news.admin.net-abuse.email; Part 3: Understanding NANAE. Spam FAQ. Archived from the original on 2004-02-12. Retrieved 2008-08-19.
- [13] "You Might Be An Anti-Spam Kook If...". Rhyolite Software. 2006-11-25. Retrieved 2007-01-05.
- [14] Scott Hazen Mueller. "What is spam?". Information about spam. Abuse.net. Retrieved 2007-01-05.
- [15] Dan Fletcher (November 2, 2009). "A Brief History of Spam". Time. Retrieved 2010-09-23.
- [16] Email metrics report, MAAWG, Nov 2010.
- [17] Valsa Raj Uchamballi, Sabyasachi Chakrabarty, Basudev Saha "An Overview of SPAM: Impact and Countermeasures", Indian Computer Emergency Response Team Enhancing Cyber Security in India, CERT-In White Paper 2005-02.
- [18] M Wong, W Schlitt, "Sender policy framework (SPF) for authorizing use of domains in e-mail, version 1" – 2006
- [19] Internet Draft J. Lyon , M. Wong "RFC 4406: Sender ID: Authenticating E-Mail" – 2006
- [20] Sourceforge, "Identified Internet Mail" <http://sourceforge.net/projects/identifiedmail>
- [21] MX Logic: http://www.mxlogic.com/news_events/01_03_05.html

Research on Evaluation Techniques for Immersive Multimedia

Aslinda Md. Hashim, Fakaruddin Fahmi Romli, Zosipha Zainal Osman
Product Quality & Reliability Engineering
MIMOS Berhad
Kuala Lumpur, Malaysia
aslinda@mimos.my, fahmi.romli@mimos.my, zosipha.zainal@mimos.my

Abstract—Nowadays Immersive Multimedia covers most usage in tremendous ways, such as healthcare/surgery, military, architecture, art, entertainment, education, business, media, sport, rehabilitation/treatment and training areas. Moreover, the significant of Immersive Multimedia to directly meet the end-users, clients and customers needs for a diversity of feature and purpose is the assembly of multiple elements that drive effective Immersive Multimedia system design, so evaluation techniques is crucial for Immersive Multimedia environments. A brief general idea of virtual environment (VE) context and ‘realism’ concept that formulate the Immersive Multimedia environments is then provided. This is followed by a concise summary of the elements of VE assessment technique that is applied in Immersive Multimedia system design, which outlines the classification space for Immersive Multimedia environments evaluation techniques and gives an overview of the types of results reported. A particular focus is placed on the implications of the Immersive Multimedia environments evaluation techniques in relation to the elements of VE assessment technique, which is the primary purpose of producing this research. The paper will then conclude with an extensive overview of the recommendations emanating from the research.

I. INTRODUCTION

Pictures can interpret a thousand and one stories compared to stories written in long sentences. This method is being utilized for thousands of years as a medium of communication. But from the point of view of equipment and techniques of drawing pictures, several changes have taken place competing with the circulation of time. Modern computer graphic has contributed a lot to the newest and potential media in order to produce pictorial communication. However, the upgrading to interactive 3D graphics system has given a large change and indirectly we are moving to new dimension.

Thus, at present it is possible to produce computer graphics images at sufficient rates that is it such images are being seen by a person where head position is automatically tracked and the viewpoint for the computer graphic will move together to follow that person’s head orientation. The technique is an ‘immersive’ technique where that particular person in an ‘immersive’ VE and being presented with computer-generated views using display panel which is a device mounted on his head to prevent sensory data for physical reality from reaching the person’s senses. This concept is being given the title or specific name as ‘virtual reality’ [1]. In 1987, one of the founder members of VPL Ins., Jaron Lanier coined the phrase

‘virtual reality’. And for the purists, immersion is virtual reality. Furthermore, Michael Heim said “virtual reality is an event or entity that is real in effect but not in fact” [2].

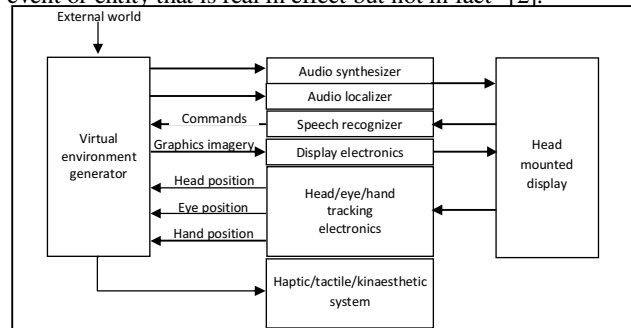


Fig. 1. Immersive virtual reality interface

Immersive Multimedia usage has been started in 1986 by Furness (1986) of the Armstrong Medical Research Laboratory at Wright Patterson Air Force Base (WPAFB). Furness reported that from that year onwards the US Air Force (USAF) has been conducting intensive research to identify future technologies whereby these technologies (Super Cockpit Concept) would meet the operational needs for the next 20 or 30 years [2]. Therefore until now, after been established for about 25 years, this technology is still being developed continuously. And nowadays we can see a lot of applications that is using virtual reality concept and purposely exist to make things easier, increase the quality of education, reduce the long term cost in industry and to be as one of the entertainment, through games.

Apart from other fields, the Immersive Multimedia system design element is essential to make ‘immersive’ VE structure explicable to users so that they effortlessly navigates and creates an attractive sense of how much is the fascinating remains to be discovered as well as establishing confidence and feelings of security within the environment [3]. Therefore, VE designers must enhance presence, immersion, and system comfort, while minimizing sickness and deleterious aftereffects. Through the development of a multi-criteria assessment technique, the current effort categorizes and integrates these VE attributes into a systematic approach to designing and evaluating VE usability. Validation exercises suggest this technique, the Multi-criteria Assessment of Usability for Virtual Environments (MAUVE) system,

This paper is sponsored by Product Quality and Reliability Engineering department of MIMOS Berhad, Kuala Lumpur, Malaysia (pqre@mimos.my)

provides a structured approach for achieving usability in VE system design and evaluation [4].

II. IMMERSIVE MULTIMEDIA: VIRTUAL WORLDS

A. Virtual Environment Context

The use of virtual environment systems is complicated because the human visual system is very sensitive to any anomalies in the imagery presented to a user. The smallest, almost imperceptible, artifact of a computer graphics system can be very apparent when motion is incorporated in the scene [2].

Computer-generated images have been created for films, generics and advertising for the past 35 years. During which time, scientific researchers, medical people and architects discovered the great potential of these images and use them to create something impossible, to visualize the invisible or simulate the non-existing. Such is the genesis of virtual worlds. Undoubtedly, these virtual worlds had two severe limitations. First, there was very little visual representation of living organisms or only very simple creatures in them and last but not least is where nobody could really enter into these worlds: the access to the virtual worlds was looking on 2D screens and 2D interaction [1].

Now, all such limitations have been overcome. New interface and 3D devices have enabled the production of a complete immersion into the virtual world. The new way of immersion into the virtual world is called Virtual Reality. Another source gives the definition that Virtual Reality is a real-space proposition involving, other than interface simulators which is a direct and real-time communication and other media representation such as audio. With the merger of the related items can represent the content Virtual Reality. Reference [1] defines Virtual Reality is the delivery to a human of the most convincing illusion possible that they are in another reality where this reality exists in digital electronic form in memory of a computer.

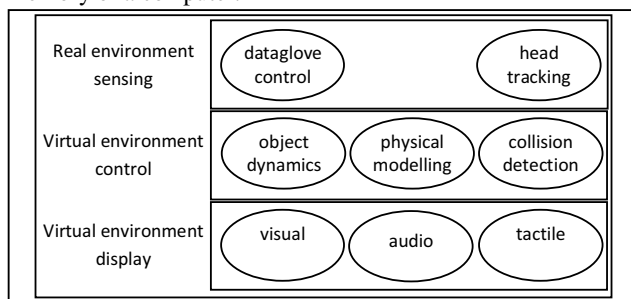


Fig. 2. Virtual Reality sub-systems

Typical Virtual Reality system containing three parts. Firstly the visual, tactile and acoustic sub-systems through which the user sees, feels and hears the virtual world. Secondly, the manual controls for navigating through the virtual world. This is a sophisticated system provides a glove containing position sensors for hand and might be a suit that covers the entire body. Finally, there are central coordinating processors and software.

Furthermore, the main human senses, vision, hearing, touch and smell will give the information on the view of the outside world. Doubled with this is the proprioceptive data where they are obtained from signals from our muscles and joints. In order to get a 'fully immersive' experience, the uses of those human sensory systems are not sufficient. The human body itself must be tracked (traces) so that movements of the human body drive changes in the action [1].

Additional factors must also be considered, such as the architecture of the virtual environment system. A poorly developed set of hardware components will lead to much inefficiency that could make the system unusable. Before even considering the synthesis of a virtual environment it is very important to understand the task requirement of the application.

B. 'Realism' Concept

Basically, realism is concerned with the fidelity of rendering of 'reality' into a particular representation. In Virtual reality this overall process can be divided into several information transformation processes. First, a part of the real world is modeled as the combination of abstract objects with certain extensions, position, colors, behavior and other attributes. Then, in the second part, these abstract objects are, driven by algorithms, rendered into a representation. A user can perceive this representation through his senses and, as the third and final transformation process, uses it to construct a mental model of the 'reality' surrounding him.

Virtual Reality is communication - An adequate approach to realism has to take into account all three-representation steps as well as their interdependencies. It is assumed that the basic idea of Virtual Reality is to mediate some information, for example one view of reality, through a medium (virtual environments) to a recipient, who is willing to get this information. That suggests the understanding of Virtual Reality as a kind of communication and comes to the conclusion that realism depends on the credibility of the representation to the user. To optimize the effectiveness of achieving this credibility one should pay attention to all phases of this communication process, i.e. the initial goals, the modeling, the options and restrictions of the medium and the abilities, experiences and expectations of the recipient, during creating and presenting virtual environments [5].

Influences on the result - The tendency of people to perceive it as an even less realistic human face is an example for the influence of perception of realism. An example for the interdependencies between modeling, the perception of the user, and his realistic impression would be the modeling of an office with a workstation on the desk. For a computer scientist as a user, the modeling of the workstation has to be relatively detailed and accurate compared to the modeling of, e.g. the window frames to create an harmonies impression, at least in the sense of the particular observer. Whereas an architect will not be interested in information about computer, but he instinctively will have a closer look on the construction details of the room. This illustrates the influence of the individual perception on the affect the representation has on the user [5].

Problem of controlling the result through design - The difficulty in designing a virtual environment is that the entire

process has to be kept in mind, while not all three steps of it can be controlled to the same extent. Only the representation as abstract objects can be influenced directly, while the perceptible representation can be influenced only indirectly by choosing and modifying the algorithm that generates it. Finally the process of perception itself is beyond the control. Additionally, in interactive real-time applications the single picture cannot be controlled totally in its impression, and there is no fixed path guiding the spectator from one point to the next. Other principles of designing and controlling the effect on the user have to be developed [5].

Lost in space or guided tours? – One important principle to represent complex information in a certain communication context is abstraction. As a part of the design process it is applied to reduce data amounts, sharpen the message by confining it to one level and modulate the information hierarchy. Thus abstraction can be employed to guide the user during the exploration in VR. The interrelation between modeling and rendering is illustrated when abstraction is used as an automated process during the VR session. However, abstraction does not apply only to modeling of the abstract objects, but to all forms of representations the user is confronted with, i.e. all effects included with the rendering of the abstract objects into a perceptible representation. The degree of abstraction might change on fixed conditions or continuously, furthermore it can depend on the distance to the observer, on the interests of the user, or the experiences the visitor of the virtual environment already made [5].

Synergy - The exploration as an entire experience – Another factor worth mentioning is the entirety of the experience in VR, stimulating multiple senses over a period of time. A coherent outlook results in more realism than detailed geometric information, physical correctness, and so on. The design should utilize all kind of simulation simultaneously. A perfectly synchronized event on multiple levels can replace data intensive elements and reinforce the user's perception. One should try to reduce the information to the necessary level, not only to get better computer performance, but more to prevent information overload and concentrate on the statements the environment includes [5].

Outlook – The introduced aspects and techniques of design control, abstraction, reduction, guidance and entirety including their interrelations and interdependencies cannot only be employed to support the realistic impression of a virtual environment. They can also be used to communicate any kind of information, be it figurative or abstract. The question before building virtual worlds including modeling, coloring, texturing, interaction etc. should be: what is the expression of this virtual environment and which conclusion from the exploration will be drawn? Furthermore, does it make sense to limit the possible expression to limitations of parts of the real world? Like any technical development the first step is to simulate existing objects, well-known situations and proved working processes. But after some time, new forms of applications, that we cannot imagine now and particular conventions for the design and the perception of virtual environment will be developed over time

just like it happened with the rise of printing. Not until this point will virtual reality unleash its full power and step out of the shadow of realism [5].

III. APPLICABILITY IN IMMERSIVE MULTIMEDIA SYSTEM DESIGN

Designing sensible and effective Immersive Multimedia environments creates an interesting design challenge for system developers and ergonomics experts. The mixture of techniques for producing visual, spatial, textual and audio representation, time-based sequencing, arrangement and manipulation of 2D images within a 3D space presents a really extra complex Immersive Multimedia experience which requires an integration of Immersive Multimedia features into a logical approach to designing VE usability.

In particular, Multi-criteria assessment of usability for virtual environments (MAUVE) can assist designers of Immersive Multimedia systems in enhancing the usability of the systems. Using this distinction, VE system interface can be characterized by the interaction techniques (as user inputs are inherently included) and multi-modal system outputs (convey interaction feedback and other system information) employed in a system design [4]. The overall MAUVE elements are shown in Figure 3.

A. Interaction Techniques

Interaction in Immersive Multimedia environment depends on various aspects. Reference [6] suggests that four of these aspects are movement, selection, manipulation and scaling. The mapping of physical motion to virtual motion is one of the most intuitive means of movement through the virtual world. Selection refers to how interaction with virtual objects (such as grabbing) requires some form of object selection, i.e., some way to indicate the target of the desired interaction. Manipulation is the specification of an object's position and/or orientation in the virtual world. This interaction can be realistic, the user grabs and moves a virtual object as he would grab and move objects in the real world, or the user can move objects in ways that have no analog in the physical world. While scaling defines as an exploration of an environment to allow a user to view some small detail by scaling up the selected object or to get a better global understanding of the environment by scaling it down and viewing it as a miniature model.

As this definition suggests, the processes involved in interaction highlights the complexities created when the interaction emphasizes the more active role which may be required whereas usability criteria associated with interaction have been herein sub-classified as: wayfinding (i.e., locating and orienting oneself in an environment); navigation (i.e., moving from one location to another in an environment); and object selection and manipulation (i.e., targeting objects within an environment to reposition, reorient and/or query)[4].

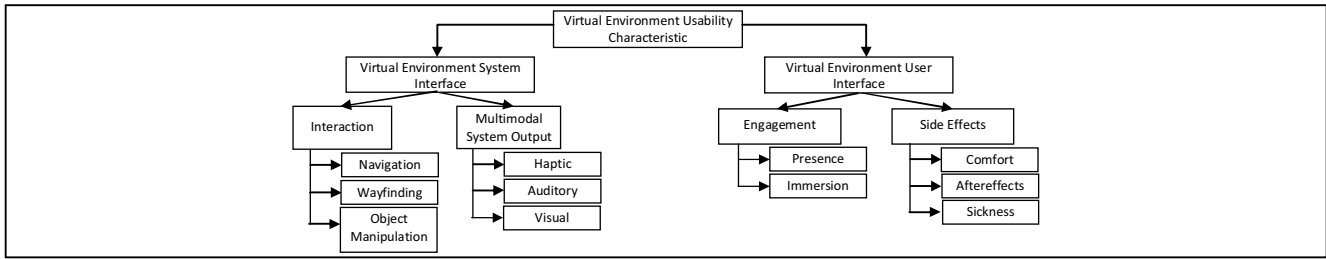


Fig. 3. Usability criteria assessed by Multi-criteria Assessment of Usability for Virtual Environment (MAUVE)

B. Multimodal System Output Techniques

A multimodal system provides a multimodal interface through which a user renders interacting with an environment more natural and intuitive by using his/her natural communication modalities, such as visual, auditory, tactile, etc. Multimodalities extend the users' usable areas while being in the dynamic locations (users can use their speech input while being in the dark place or the users can provide a gesture or touch-based input to the system while being in the noisy environment). Whereas audio-based feedback might be more effective when the users are doing some outdoor activities like playing football, drawing, walking etc. The haptic modality (vibration, pressure, temperature etc.) might be effective when the users are in the noisy situation like in railway station, bus stop or in airport [7].

These modalities are used to engage human perceptual, cognitive, and communication skills in understanding what is being presented in a virtual world [8]. While interacting with multimodal-equipped system, user should recognize about the character of the system, like how many ways to interact, language to communicate, error prevention techniques, and how to get a feedback from it, in case the user gets lost [7].

Anyhow, in ensuring multimodal system design to meet system requirements and allows users to interact with a VE free of frustration and irritation is thus of paramount importance. An underlying principle for such guidelines is that any form of information (i.e., visual, auditory, haptic) presented to users should be readily understood, unambiguous, and necessary to complete required tasks [4].

IV. EVALUATION TECHNIQUES IN IMMERSIVE MULTIMEDIA

A. Approach

The research also identifies the established evaluation techniques based on a recognized task analysis and categorization of techniques applying a taxonomy of techniques for the universal Immersive Multimedia system design elements, by which could evaluate interaction and multimodal system output techniques.

One way to verify the generality of the taxonomies is through the process of categorization. If an existing techniques for the task fit well into the taxonomy that can ensure more of its correctness and completeness. Categorization also serves as an aid for evaluation of techniques. Fitting techniques into a taxonomy makes explicit their fundamental differences, and able to determine the effect of choices in a more fine-grained manner [9]. Literature presents various taxonomies that can be

implemented and it appears that strategies have been successfully established to facilitate Immersive Multimedia system evaluation.

Travel, which means the control of user viewpoint motion through a Immersive Multimedia, is an important and universal user interface task which needs to be better understood and implemented in order to maximize users' comfort and productivity in Immersive Multimedia systems. The distinguish travels from navigation or wayfinding, which refer to the process of determining a path through an environment to reach a goal. The work attempts to comprehend and categorize the techniques that have been proposed and implemented, and to demonstrate an experimental method which may be used to evaluate the effectiveness of travel techniques in a structured and logical way [10].

Figure 4 shows high-level entries in virtual travel techniques taxonomy. There are three components in a travel technique, each of which corresponds to a design decision that must be made by the implementor. *Direction/Target Selection* refers to the method by which the user "steers" the direction of travel, or selects the goal position of the movement. *Velocity/Acceleration Selection* methods allow the user/system to set speed and/or acceleration. Finally, *Input Conditions* are the way in which the user or system specifies the beginning of time, duration, and end time of the travel motion [10].

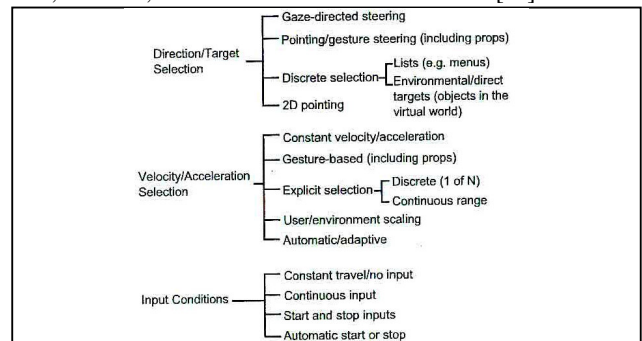


Fig. 4. Taxonomy of virtual travel techniques

In fact, manipulation/selection of objects in virtual environments is often awkward and inconvenient. A lack of a tactile feedback, tracker noise, poor design of interaction techniques, and other factors can make the simple task of grabbing and moving a virtual object a frustrating experience. Similarly, development of effective VR applications also requires comprehensive understanding of immersive manipulation/selection and, in particular, which virtual tools and techniques should be used and how they should be

designed for effective and ease of use [11, 12]. As illustrated, figure 5 shows a taxonomy for the interested tasks of selection and manipulation of interaction techniques.

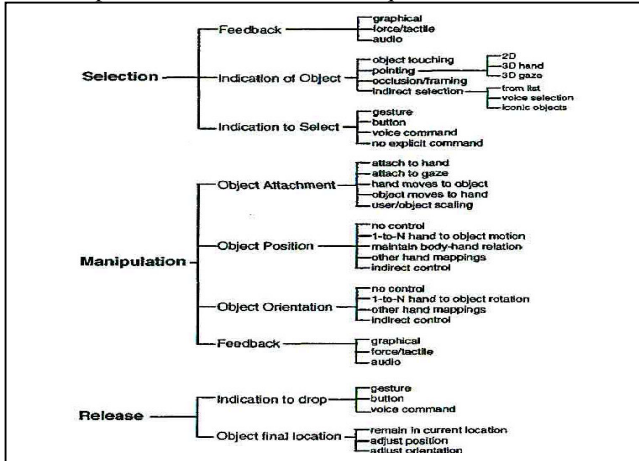


Fig. 5. Taxonomy of selection/manipulation techniques

Nevertheless, with multi-sensory interfaces the user can potentially perceive and assimilate multi-attributed information more effectively. By mapping different attributes of the data to different senses, such as the visual, auditory and haptic sense, it may be possible to better understand large data sets [13].

In relation to this Keith V. Nesbit introduces a classification of the multi-sensory design space called the Multi Sensory (MS)-Taxonomy. The classification makes a primary division of the design space using the types of metaphors used in information displays. This results in three main classes, *Spatial metaphors*, *Direct metaphors* and *Temporal metaphors* [13].

Spatial metaphors concern the way pictures, sounds and forces are organised in space and can be described for the visual, auditory and haptic senses. Thus different types of spatial metaphors may be described for each sense. *Spatial visual metaphors* concern the way pictures are organised in space. *Spatial auditory metaphors* concern the way sounds are organised in space. *Spatial haptic metaphors* concern the way forces are organised in space [13].

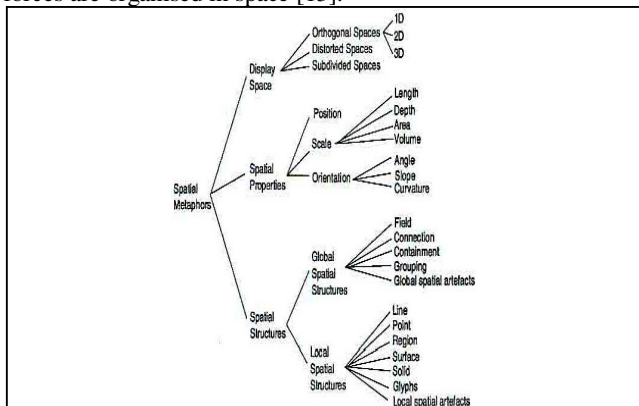


Fig. 6. General concepts that describe Spatial Metaphores

Meanwhile *Direct metaphors* are concerned with direct mappings between sensory properties and some abstract information. For example, a specific colour, the volume of sound or the hardness of a surface may be used to represent a particular data attribute. Once again, a class of direct metaphors can be defined for each sense. *Direct visual metaphors* concern the perceived properties of pictures. *Direct auditory metaphors* concern the perceived properties of sounds. *Direct haptic metaphors* concern the perceived properties of touch [13].

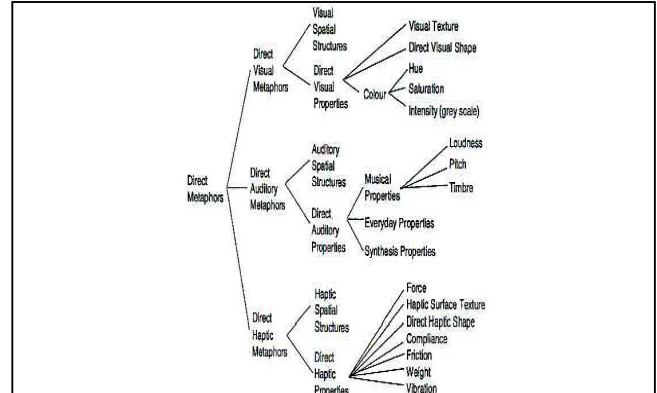


Fig. 7. General concepts that describe Direct Metaphores

On the other hand *Temporal metaphors* are concerned with how we perceive changes to pictures, sounds and forces over time. The emphasis is on displaying information by using the fluctuations that occur over time. Once again Temporal metaphors can be considered for each of the senses. *Temporal visual metaphors* concern the way pictures change with time. *Temporal auditory metaphors* concern the way sounds change with time. *Temporal haptic metaphors* concern the way haptic stimuli change with time [13].

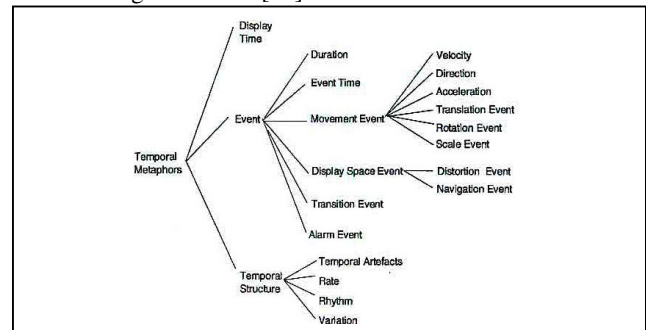


Fig. 8. General concepts that describe Temporal Metaphores

The major strength of those hypotheses is that it considers for the diversity in ideas and focuses among various ways of interaction and multimodel system output evaluation techniques in the Immersive Multimedia system. In particular, it provides an excellent clarification for the taxonomy, by justifying the rational of taxonomy tasks and subsequent tasks.

B. Classification of Evaluation Techniques in Immersive Multimedia

The design of interaction techniques and user interfaces for virtual environments must be done with extreme care in order to produce useful and usable systems [10].

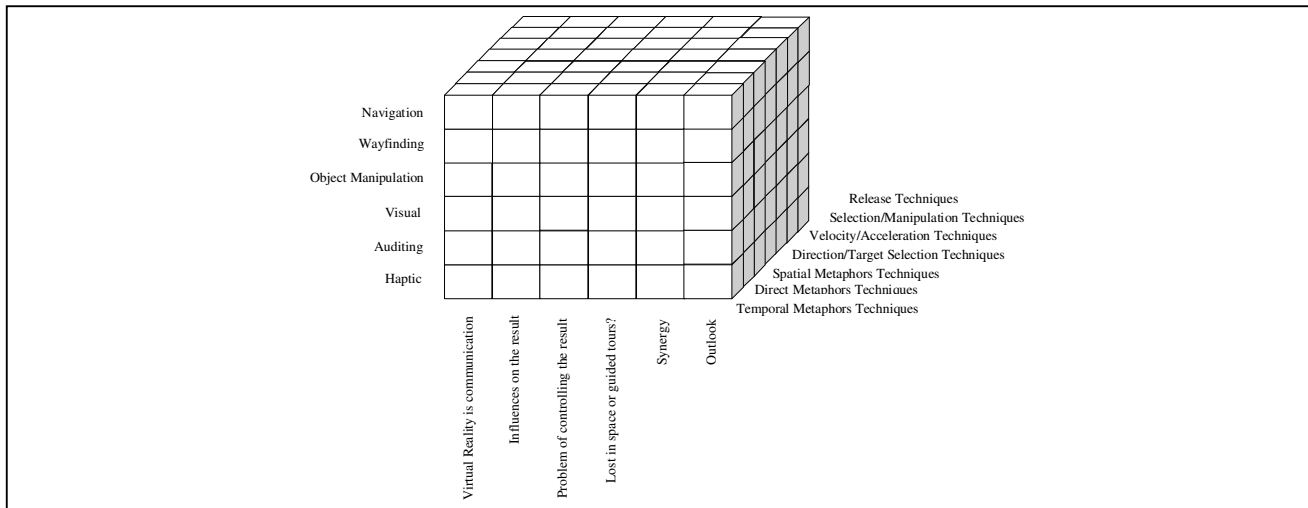


Fig. 9. Cubic Immersive Multimedia Evaluation Facet

The citation enclosed in the proposal shows that Immersive Multimedia evaluation is something more than to produce and deliver with accomplish the purposes, contents, structures and process of realization of the end-users, clients and customers needs. It is more than individual exploration and contents. It is more than different perception and activities in and beyond the non-virtual environment. It grasps all the elements stated above. It can be considered as assurance approach to find design issues of complex information in a communication context to be exchanged in abstraction manner of virtual environment. Hence, Immersive Multimedia evaluation is a multidimensional evaluation concept.

Immersive Multimedia evaluation should meditate various quality factors that would be essential in evaluation techniques and strategies that might form the integral part of the Immersive Multimedia evaluation model. It is sensible and significant to apply the multidimensional evaluation concept in the Immersive Multimedia evaluation techniques. The authors proposed the model of Cubic Immersive Multimedia Evaluation Facet aiming the future needs of diversity of feature and purpose of Immersive Multimedia system.

The Cubic Immersive Multimedia Evaluation Facet is represent as a cube having three characteristic dimensions of system usability measure, evaluation techniques and realism measure where influenced techniques and different strategies of evaluation which can be applied. The cube is an ideal three-dimensional mold which can assist us to have a methodical view at what the coverage of Immersive Multimedia evaluation techniques and evaluation strategies is.

The first dimension of the cubic evaluation highlights the system usability measure as suggested by [4]. This dimension is a usability evaluations can be used for; finding usability issues in a design; validating the usability of a design; providing baseline metrics to measure future progress; and producing analytical approaches to determine which changes to a design will have the greatest impact on its usability [4].

The second dimension highlights the realism measure that across over the system usability measure dimension. The increasing complexity of the dynamic and interactive

representation of the world to previous pictorial techniques: on the one hand, the projection plane can be moved at through space, and consequently the potential perspectives are infinite; on the other hand, the programming enables the environment to react to the user's actions [14]. It's essential to overview across the many system usability measure in order the practical evaluators could contemplate how effective the interaction with the object from the virtual environment, the manipulation, the feeling of the human user towards virtual environment as well as 'realism' context.

The third dimension highlights evaluation techniques which practical evaluators must be able to apply evaluation approaches correctly—the scientific aspect of program evaluation, beforehand they must be able to select the evaluation approach that complements the needs and realities they face—the art aspect of program evaluation. The evaluation taxonomy is enhancing understanding of evaluation's art aspect. In fact, taxonomy can assist with conceptualization of evaluation needs, focusing of evaluation activities, and identification of available evaluation means that suit a program's needs and realities [15].

In particular, these characteristic dimensions identify those techniques that are applied in a generic context and an element-specific context. The context of Immersive Multimedia evaluation naturally enforces improvement on the applicability and generality of results. Hence, results of evaluations performed in a generic context can typically be applied more generally than results of an element-specific evaluation technique, which may produces qualitative or quantitative types of results reported.

The types of results reported are not designed to be mutually exclusive, and are instead designed to convey Immersive Multimedia evaluation techniques. For instance, particular Immersive Multimedia evaluation techniques may produce both quantitative and qualitative results. Certainly, many of the identified techniques are adaptable enough to provide awareness at many levels.

The intention of the cubic evaluation is to unite the possible of Immersive Multimedia environment behavior in the mutual

experiences of human user and to enhance the evaluation techniques to incorporate non-virtual environment and virtual environment aspect in order to produce virtual world in reality world.

ACKNOWLEDGMENT

Thanks to the anonymous reviewers for valuable remarks and suggestions.

REFERENCES

- [1] Harrison, David & Jaques, Mark., "Experiments in Virtual Reality". Butterworth Heinemann, 1996.
- [2] Kalawsky, Roy S., "The Science of Virtual Reality and Virtual Environments". Addison-Wesley Publishing Company, 1993.
- [3] Bell, B., Greene, T., Fisher, J. and Baum, A., "Environmental Psychology". Orlando: Harcourt Brace College Publishers, 1996
- [4] Kay M. Stanney, Mansooreh Mollaghasemi, Leah Reeves, Robert Breaux, David A. Graeber, "Usability engineering of virtual environments (VEs): identifying multiple criteria that drive effective VE system design", International Journal of Human-Computer Studies, Volume 58, Issue 4, April 2003, Pages 447-481, ISSN 1071-5819, 10.1016/S1071-5819(03)00015-6.
- [5] P. Astheimer, F. Dai, M. Göbel, R. Kruse, S. Müller, G. Zachmann, "Realism in virtual reality". Artificial Life and Virtual Reality, John Wiley & Sons, 1994.
- [6] Mark R. Mine, "Virtual Environment Interaction Techniques", University of North Carolina, Chapel Hill, NC. TR95-018, 1995.
- [7] Dilip Roy, "Distributed Multimodal Interaction in a Smart Home Environment", Umea University Department of Computing Science Umea Sweden, 2009.
- [8] Turk & Robertson, "Perceptual user interfaces", Communications of the ACM, 43 (3) (2000), pp. 33-34, 2000.
- [9] Doug A.Bowman, "Interaction Techniques for Immersive Virtual Environments: Design, Evaluation, and Application", Graphics, Visualization, and Usability Center College of Computing Georgia Institute of Technology, 1998.
- [10] Doug A.Bowman, Donald B. Johnson, Larry F. Hodges, "Testbed Evaluation of Virtual Environment Interaction Techniques", Graphics, Visualization, and Usability Center College of Computing Georgia Institute of Technology, 1999.
- [11] Mine, M., "Virtual environment interaction techniques.", UNC Chapel Hill Computer Science Tech. Report TR95-018, 1995.
- [12] Herndon, K., van Dam, A., Gleicher, M., "The challenges of 3D interaction: a CHI'94 workshop.", SIGCHI Bulletin, Vol. 26, N 4, 1994.
- [13] Keith V. Nesbitt, "MS-Taxonomy: A Conceptual Framework for Designing Multi-sensory Displays", Charles Sturt University, School of Information Technology, Bathurst, Australia, 2004.
- [14] Laia Pujol-Tost, "Realism in Virtual Reality applications for Cultural Heritage", The International Journal of Virtual Reality, 2011, 10(3): 41-49.
- [15] Huey-Tsyh Chen, "Practical Program Evaluation Assessing and Improving Planning, Implementation, and Effectiveness", SAGE Publications, Inc, 2004.

Collision Avatar (CA): Adding collision objects for human body in augmented reality using Kinect

Kairat Aitpayev

Computer Science Department
Center for Energy Research and University of Technology
of Belfort-Montbéliard
Astana, Kazakhstan and Belford, France
kairat.aitpayev@nu.edu.kz

Jaafar Gaber

Computer Science Department
University of Technology of Belfort-Montbéliard
Belford, France
gaber@utbm.fr

Abstract— The purpose of this study is the improvement of real and virtual world interaction in real-time by uses of Kinect. We propose adding geometric shapes as a collision object for human body parts in augmented reality with real-time animations. In this work the concept of collision Avatar (CA) is presented. We will describe different methods which creates geometric shapes, scale them according to body part sizes. Also we compare several types of real-time animations using different methods.

Index Terms— Augmented Reality, Motion Capture, Virtual Character, Ubiquitous Computing.

INTRODUCTION

Development of ubiquitous computing is popular topic in the recent times. Augmented Reality is a subtopic in the ubiquitous computing domain, which is composed of real and virtual worlds.

Because computer performance is continually growing, we can update augmented reality with Kinect's information about the human body. It is possible to apply physical libraries to virtual objects and add interaction between them. Core technologies of augmented reality and Kinect are position recognition, image recognition, orientation recognition and motion capture. By merging these two technologies we can increase functionality and uses in game development, mental treatment, marketing, etc. Using modern hardware, we can process these data using mobile phones as well as notebooks.

The most similar related work to ours is Augmented Mirror[3]. It is using special motion capture (MoCap) system which consists of Kinect, Wii and gyroscope. The idea of this work is to animate augmented 3D character and other objects in real-time. The Mocap system tracks motions of human and apply them to character. So that other people can interact with this cartoon character as with real person. In our approach we are changing the idea of interactions between humans and augmented reality. Instead of using 3D avatars we are giving ability to human to use own body for interaction with augmented object.

Our approach is based on making the human body a physical part of augmented reality and add the ability for interactions with other augmented objects. Previously, it was necessary to wear a special suit with infrared LEDs, or attach

markers. The development of depth cameras like Microsoft Kinect removes the need of these additional equipment. With Kinect we can measure and track users body and create CA and animate them in real-time.

A CA which works with augmented reality in real-time, proves of this concept. When it was developed, the applications efficiency was emphasized by using Kinect sdk for Windows, C# libraries "Goblin XNA", "DigitalRune" and 3DCG software "3ds Max".

BODY PARTS CONSTRUCTION

For body construction four different types of shapes are used. These are sphere, capsule, box and the last being the convexShape¹ object. Depending on the application, different

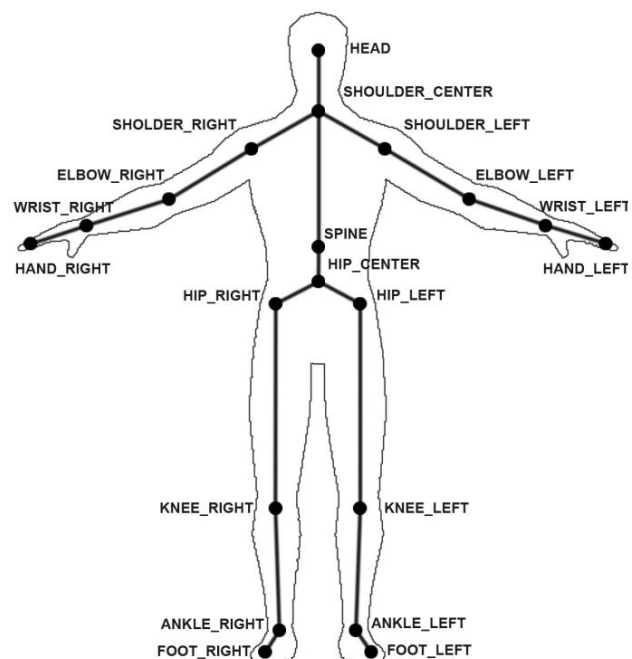


Fig. 1. Skeleton structure, which consists of 20 joints.

¹ Strapping a film around pointed shapes to construct a new one. Digital Rune library has such shape.

types of shapes can be applied to represent the human body in augmented reality. This section will describe the detailed construction of the virtual body. Actually, according to applicable tasks, the virtual body can be simplified for management of computing resources. As it will be explained in the section below.

A. Head and Neck

For the human head the most suitable shape is the sphere. To create a sphere only one parameter is needed, the diameter or the radius. Logically, the height of head is needed for the diameter so that the sphere would cover the entire head. The problem of calculating this height, is that the head should be ideally be detached from rest of the body on the depth image, but this would require additional calculations. Instead of calculating this height, it is much easier to calculate the width at the point of Head joint from the Kinect skeleton. Then using standard head proportions to convert this width into proportionate head heights as illustrated on Fig.2a. Since all tests were made on men with short hairstyles, there may be some problems with other people, especially females with long hair. As a solution for this problem, we can get the height of the head by measuring the proportions between the arms and head as illustrated on Fig.3. The neck is representing by capsule shape. Distance between the Head and ShoulderCenter

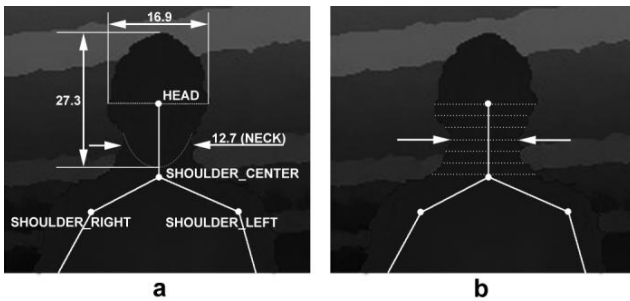


Fig. 2. Left image shows Head and Neck proportion for first method. Right image shows searching region for neck width and which line should be taken.

joints are taken as the length of this capsule. For finding the width we propose two methods. First, the easiest method is to use proportions from Fig.3. The second method for a more accurate construction would find the thinnest place along the neck bone and take this as the width as illustrated on Fig.2b.

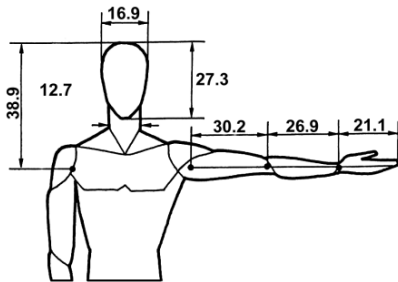


Fig. 3. Upper body proportions.

B. Arms and Legs

Construction of augmented object for hands and legs is the same as illustrated on Fig.3. Capsules are representing the following bones: shoulder - elbow - wrist, hip(right/left) - knee - ankle. To find the dimensions of the capsules length and radius the same technique as with the neck object were applied. To simplify counting of pixel, it is better to use pose detection[1] as illustrated on Front view of Fig.4. In these positions, bone length calculations are more accurate and it is enough to calculate horizontal or vertical pixels at the center of bones, instead of applying trigonometric calculations to find pixels which lie perpendicular to the bones. For representing hands and feet the most suitable shape is a box. To calculate all dimensions of a box, the length is taken from the bone, following proportions for the length:width:height are taken: for the arms - 5:3:2; legs - 4:3:2 respectively.

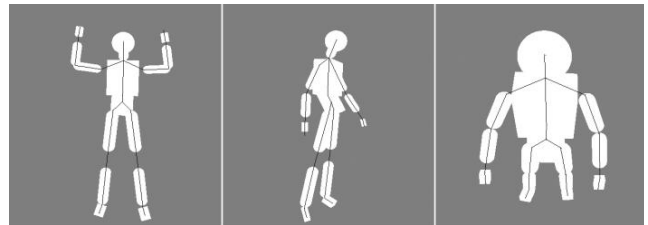


Fig. 4. Collision object from front, left and top views. Front view represents the most suitable position for calculating object's dimensions.

C. Torso and Pelvis

For the pelvis we are using flattened sphere with ratio between diameter and height as 2:1. The distance between HipLeft and HipRight joints were multiplied by two and taken as the diameter of the sphere.

The most complex augmented object of the body is the torso, because in real life it could have very different types of shapes, that's why to create convexShape three separate objects is taken. Two of them are capsules, top capsule represents the shoulders, bottom represents pelvis and sphere is needed to manage the belly size as illustrated on Fig.5. Top capsule's length RE taken from the distance between ShoulderLeft and ShoulderRight joints and the bottom capsule between HipLeft and HipRight then multiplied by two. The Radius for these two capsules calculated by ratio length:radius - 5:1.

Now to create the belly; only if it is needed, next method could be applied. Difference between distances d1 and d2 at ShoulderCenter and HipCenter joints on the depth image should be scaled according to the virtual environment and then taken as the radius of sphere. After this sphere is situated as illustrated on Fig.5.

This is an accurate creation of the front part of a torso object which will almost fits to any real human body and could be used for example to wear virtual clothes in such applications like virtual fitting room. Usually such complex creation is not required and convexShape objects can be replaced by a capsule or box.

There are several methods for animating collision objects in real-time. This paper explains two widely used methods. The first method is easy to use and implement, because movements are carried out by positioning not by angles. This method needs a physical library which has a ragdoll² and IK-Solvers³[2]. The second method calculates the rotation of each bone about three axes and apply them to related collision object.

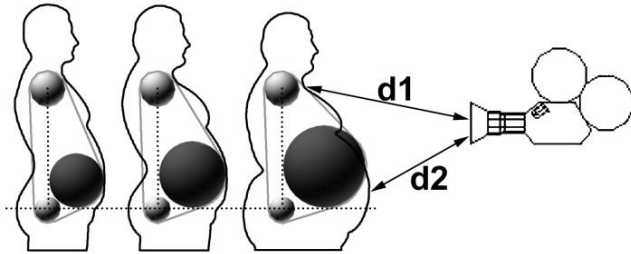


Fig. 5. Construction of Torso. Shape positions and calculations.

A. Ragdoll animation

After creating all collision objects, they will be joined together as a ragdoll. IK-Solvers are applied for arms and legs. For this method, Kinect's skeleton should be scaled to match the ragdoll's size in the 3D environment. Initially the whole ragdoll is positioning and rotated by coordinates of HipCenter and Spine joints. For animating arms and legs its needed to apply coordinates of WristLeft, WristRight, AnkleLeft and AnkleRight to related IK-Solvers. Note that hand and foot joints are not used, instead of wrists and ankles. This is because Kinect provides pure recognition of those joints and it's more accurate to use IK-Solvers between two bones than three. In applying this method the head, hands and feets should be fixed according to the parent bones otherwise they would not have natural, free movements according to physical parameters. Positions of elbows and knees are calculated by IK-Solvers so they are uncontrolled by the user's movements. Since this method is based on positioning it would be hard to rotate the ragdoll's body about the vertical axis. This is because Kinect provides only x, y, z coordinates of joints, and not the rotation angles of bones.

This method is easier in implementation by using features of the physical library, but it has some disadvantages which will be removed in the second method below.

B. Animation by angles

This method has big advantages since it reflects exactly all movements from user to CA and additionally, can be applied to avatars which have different proportions of bones according to the Kinect's skeleton. The method is based on the calculation of

² Type of procedural animation that is often used as a replacement for traditional static death animations in video games and animated films.

³ Inverse kinematics is important to game programming and 3D animation, where it is used to connect game characters physically to the virtual world.

rotation matrices for each bone according to the global coordinate system.

In our implementation we used Euler angles and matrices created through the standard method createLookAt(). Euler angles is more understandable to humans and it is better to use during debugging. After calculation of all Euler angles they must be combined together to quaternion or rotation matrix. Nowadays a lot of 3D libraries such as XNA have special class methods for these purposes, for example [Matrix, Quaternion].CreateFromYawPitchRoll(). By applying Euler angles separately they work fine, but while joining them together they cause a problem, because of the rule of cosine function: $\cos(\text{angle}) = \cos(-\text{angle})$ in the third and fourth quarters of a unit circle they give wrong orientation. It means that between 180-360 degree, the rotations of collision objects will be different from the human body.

CreateLookAt() method gives good results and is easy to use. Each bone consist of two joints; the root joint will be used as position point and the child joint as look at point. This method calculates rotations about two axes, the rotation about the third axis must be calculated by additional ways or set up as constant. For capsule and sphere objects, rotation about the third axis is not important since they have rounded shapes. Box objects which are used for hands and foots can have constant rotation. Body objects needs additional calculations for the angle about the vertical axis. This angle can be calculated between ShoulderRight and ShoulderLeft joints and multiplied by rotation matrix of the torso object. If pelvis is represented by a box, the same method can be applied to it, but now using HipLeft and HipRight joints.

C. Comparison of the two methods

The first method is easier in implementation, but has limited control of 3D avatars. It can be applied for collision objects (ragdoll) as well as for the rigged 3D character. Weaknesses of this method can be improved by merging it with the second method for rotating torso and pelvis about three axes.

The second method is a little harder in implementation and needs more computing resources, because of the calculations of angles in each frame. In comparison to the first method, this method have a more accurate real-time animation. This method can be applied only for collision objects, since a skeleton of rigged character needs angles of bones relative to their parent bones, but not to the global coordinate system.

Since the second method gives more accurate result, we have used it for our implementation and testing.

D. Improvements of Kinect information relevance

Information obtained from Kinect about position of the user's body were not always accurate. This caused the length of bones to change depending on position. To make the problem less noticeable, several solutions are offered:

a. Rescaling collision objects in each frame, of course this will strongly effect to performance and does not guarantee best result.

b. Set parameters of collision objects manually taking to consideration different variants of positions. This will make very accurate construction, but will be intended only for people with same body proportions.

c. The uses of body proportions and then apply them to calculate the lengths and sizes of collision objects. This approach needs additional gender and age recognition algorithms. Taking to consideration that not all people have a same proportions of body it will not give a perfect result, but at least would give some improvements.

CONCLUSION AND FUTURE WORK

Since we used only one Kinect motion capturing device, this was not accurate enough and we had some problems with the measurement of bones, especially when the user rotated 90 degrees from the camera. New Kinect sdk for Windows has better motion capturing functions and can support upto 6 Kinect cameras.

By using this sdk, results can be improved by hardware or software means. By hardware, we mean usage of two and more cameras which will increase the freedom of motion and skeleton recognition. This solution is more suitable for animation studios with appropriate space and budget. We however, are primarily focussed on solutions which could be used by end users with only one Kinect. Because of this, our approach is more based on software means. Alone, Kinect sdk updates often and there are big improvements in the latest version update compared to its first release. From our side we

are developing algorithms which will predict positions of invisible joints and increase the freedom of user's motion, which Kinect has yet to develop.

ACKNOWLEDGMENTS

This work is supplied by the EU Erasmus Mundus Action 2 TARGET program (EM A2) for fundamental exchange research.

REFERENCES

- [1] Jarrett Webb and James Ashley, *Beginning Kinect Programming with the Kinect SDK*, 1st ed. ISBN-13:978-1430241041, Apress, 2012.
- [2] DigitalRune, *3D game engine for the Microsoft .NET Framework and the Microsoft XNA Game Studio*, <http://www.digitalrune.com/>, 2006-2011.
- [3] Hiromitsu Sato and Michael Cohen, *Using Motion Capture for Real-time Augmented Reality Scene*, HC '10 Proceedings of the 13th International Conference on Humans and Computers, 2010.
- [4] Yichen Tang, Billy Lam, Ian Stavness and Sidney Fels, *Kinect-based Augmented Reality with Perspective Correction*, SIGGRAPH, 2011.
- [5] Lucia Vera, Jesus Gimeno, Inmaculada Coma and Marcos Fernandez, *Augmented Mirror: Interactive Augmented Reality System Based on Kinect*, International Federation for Information Processing, 2011.

A Comparative Study on Feature Selection in Chinese Spam Filtering

Yan Xu^{1,2}

1. Beijing Language And Culture University, Beijing, China

2. Institute of Computing Technology, Chinese Academy of Sciences
xuy@blcu.edu.cn

Abstract

Feature selection plays an important role in Spam Filtering. Automatic feature selection methods such as document frequency thresholding (DF), information gain (IG), and so on are commonly applied in spam filtering. Spam filtering can also be seen as a special two-class text categorization (TC) problem. Many existing experiments show IG is one of the most effective methods in text categorization task. However, what is the most effective method on spam filtering? As we all know there was not a systematic research about these feature selection methods on spam filtering. This paper is a comparative study of feature selection methods in spam filtering. The focus is on aggressive dimensionality reduction. We explore 2 classifiers (Naïve Bayes and SVM), and run our experiments on Chinese-spam collection. Six methods were evaluated, including term selection based on document frequency (DF), information gain(IG), χ^2 feature selection method, expected cross entropy (ECE), the weight of evidence for text (WET) and odds ratio (ODD). We found ODD and WET most effective in our experiments. In contrast, IG and χ^2 had relatively poor performance due to their bias towards favoring rare terms, and its sensitivity to probability estimation errors.

Keywords: artificial intelligence; spam filtering; feature selection, text classification, DF

1. Introduction

As we all know that the increasing volume of spam has become a serious threat not only to the Internet, but also to society. For the business and educational environment, spam has become a security issue. Spam has gone from just being annoying to being expensive and risky. Spam is commonly defined as unsolicited email messages. Since May 2003 the amount of spam exceeded legitimate emails [11]. From China's anti-spam alliance data show that: in July 2011, spam accounted for an average total 89% [12]. Of the total e-mail on average 10 e-mails, only one is a normal e-mail,

and the rest are all spam. In April of Symantec's report, China's top ten spam ranked seventh in the country of origin[13], the report from Anti-spam center of ISC showed that the spam in Chinese take a proportion of 21.6%[14]. So we can see that how important for us to find a competitive Method to filter spam, especially Chinese spam. So, how to find an efficient spam filtering methods has always been a concerned problem by all sectors of society.

The goal of spam categorization is to distinguish between spam and legitimate email messages. A growing number of statistical classification methods and machine learning techniques have been applied to spam filtering in recent years, including multivariate regression models, nearest neighbor classification, Bayes probabilistic approaches, decision trees, neural networks, symbolic rule learning and inductive learning algorithms. Many researches in spam filtering have been centered on the more sophisticated classifier-related issues. Bayesian classifiers are the most widely used method in this field. The representative figure is Sahami et al [15] and Greek scholar Androutsopoulos[16]. Sahami et al trained a Naïve Bayesian classifier on manually categorized legitimate and spare messages, reporting impressive performance on unseen messages. Other machine learning methods, including Support Vector Machine (SVM)[17], Rocchio, kNN and Boosting[18], have also been applied in anti-spam filtering. In addition, Gordon V. Cormack has proposed the Dynamic Markov Compress Model for spam filtering task in TREC and is proven effective[19]. Nevertheless, to our best know there was few research on a key technology in text classification – feature selection.

As we all know, a major characteristic, or difficulty, of spam filtering problems is the high dimensionality of the feature space. Automatic feature selection methods are designed to solve this problem, which include the removal of non-informative terms according to corpus statistics, and the construction of new features which combine lower level features(i.e., terms) into higher level orthogonal dimensions. Lewis & Riguette used an information gain measure to

aggressively reduce the document vocabulary in a naïve Bayes model and a decision-tree approach to binary classification. Wiener et al. used a χ^2 statistic to select features for input to neural networks. Yang and Schutze et al. used principal component analysis to find orthogonal dimensions in the vector space of documents. Yang & Wilbur used document clustering techniques to estimate probabilistic “term strength”, and used it to reduce the variables in linear regression and nearest neighbor classification.

In recent years, a growing number of statistical classification methods and machine learning techniques have been applied in this field. Many feature selection methods such as document frequency thresholding (DF), information gain measure (IG), mutual information measure (MI), and so on have been widely used. Many existing experiments show IG is one of the most effective methods [1][2][3].

But, in the task of spam filtering, how will popular feature selection methods perform in the spam filtering task? Which is the best feature selection algorithm? Will a state-of-art feature selection method also performs outstanding in spam filtering?

The focus in this paper is the evaluation and comparison of feature selection methods in the reduction of a high dimensional feature space in spam filtering task. We use 2 classifiers and employ the 6 feature selection methods in our corpora, 2005-Jun. the 6 methods we explored in our experiments are: Information Gain (IG), Document Frequency (DF), χ^2 statistic (CHI), Expected cross entropy (ECE), Weight of evidence for text(WET), Odds ratio(ODD).

We design a series experiments to compare different feature selection methods. The experiments in public e-mail corpus show an effective result ODD and WET algorithms are very competitive feature selection method for anti-spam filtering tasks, especially for Chinese spam filtering.

Section 2 describes the term selection methods. Due to space limitations, we will not include phrase selection and approaches based on principal component analysis. Section 3 describes the classifiers and the document corpus chosen for empirical validation. Section 4 presents the experiments and the results. Section 5 discusses the major findings. Section 6 summarizes the conclusions.

2. Related Work

Feature selection is an important step in spam filtering. In recent years, a growing number of statistical classification methods and machine learning

techniques have been applied in this field. The prevailing feature selection methods such as Information Gain (IG), Document Frequency (DF), χ^2 statistic (CHI), Expected cross entropy (ECE), Weight of evidence for text (WET), Odds ratio (ODD). They are all explored in our experiment, and they will be introduced respectively as follow:

2.1 Document frequency thresholding

Document frequency is the number of documents in which a term occurs. Only the terms that occur in a higher number of documents are retained. We computed the document frequency for each unique term in the training corpus and removed from the feature space those terms whose document frequency was less than some predetermined threshold. The basic assumption is that rare terms are either non-informative for category predictions, or not influential in global performance. In either case removal of rare terms reduces the dimensionality of the feature space. Improvement in categorization accuracy is also possible if rare terms happen to be noise terms.

DF thresholding is the simplest technique for vocabulary reduction. It easily scales to very large corpora with a computational complexity approximately linear in the number of training documents. However, it is usually considered an ad hoc approach to improve efficiency, not a principled criterion for selection predictive features. Also, DF is typically not used for aggressive term removal because of a widely received. Assumption in information retrieval. That is, low-DF terms are assumed to be relatively informative and therefore should not be removed aggressively. We will re-examine this assumption with respect to spam filtering task.

2.2 Information gain

Information gain is commonly used as a term goodness criterion in machine learning [1]. It measures the amount of information obtained for category prediction by knowing the presence or absence of a term in a document. Let $m \{c_i\}_{i=1}^m$ denote the set of categories in the target space. The information gain of term t is defined to be:

$$G(t) = -\sum_{i=1}^m p(c_i) \log p(c_i) \\ + p(t) \sum_{i=1}^m p(c_i | t) \log p(c_i | t) \\ + p(\bar{t}) \sum_{i=1}^m p(c_i | \bar{t}) \log p(c_i | \bar{t})$$

Given a training corpus, for each unique term the information gain is computed and those terms whose information gain is less than some predetermined threshold are removed from the feature space.

2.3 χ^2 statistic (CHI)

The χ^2 statistic measures the lack of independence between t and c and can be compared to the χ^2 distribution with one degree of freedom to judge extremeness. Using the two-way contingency table of a term t and a category c , where A is the number of times t and c co-occur, B is the number of time the t occurs, and N is the total number of documents, the term-goodness measure is defined to be:

$$\chi^2(t, c) = \frac{N * (AD - CB)^2}{(A + C) * (B + D) * (A + B) * (C + D)}$$

A 、 B 、 C 、 D represent quantity of document, showing in the following table, $N = A + B + C + D$

	type c_i	not type c_i
t included	A	B
t not include	C	D

The χ^2 statistic has a natural value of zero if t and c are independent. We computed for each category the χ^2 statistic between each unique term in a training corpus and that category, and then combined the category-specific scores of each term into two scores:

$$\chi_{avg}^2(t) = \sum_{i=1}^m P_r(c_i) \chi^2(t, c_i)$$

$$\chi_{max}^2(t) = \max_{i=1}^m \{\chi^2(t, c_i)\}$$

The computation of CHI scores has a quadratic complexity, similar to IG, and χ^2 values are comparable across terms for the same category. However, this normalization breaks down (can no longer be accurately compared to the χ^2 distribution) if any cell in the contingency table is lightly populated, which is the case for low frequency terms. Hence, the χ^2 statistic is known not to be reliable for low-frequency terms.

2.4 Expected cross entropy (ECE):

$$ECE(t) = p(t) \sum_{i=1}^n p(c_i | t) \cdot \log_2 \frac{p(c_i | t)}{p(c_i)}$$

The only difference from Information gain is that expectations cross entropy method do not take words which do not exist.

2.5 Weight of evidence for text(WET):

$$\begin{aligned} WET(t) &= p(t) \cdot \sum_{i=1}^n p(c_i) \cdot \left| \log_2 \frac{Odds(c_i | t)}{Odds(c_i)} \right| \\ &= p(t) \cdot \sum_{i=1}^n p(c_i) \cdot \left| \log_2 \frac{p(c_i | t)(1 - p(c_i))}{p(c_i)(1 - p(c_i | t))} \right| \end{aligned}$$

In this function,

$$Odds(c_i | t) = \frac{p(c_i | t)}{1 - p(c_i | t)} \quad \text{and} \quad Odds(c_i) = \frac{p(c_i)}{1 - p(c_i)}$$

is the evaluation function, and the later is a relative new evaluation function. It measures the difference between the probability of class and conditional probability of class for items, and only considers the t happen in the text.

2.6 Odds ratio(ODD):

It is designed for two-class classifier, which is defined as follow:

$$\begin{aligned} OR(t) &= \log_2 \frac{Odds(t | pos)}{Odds(t | neg)} \\ &= \log_2 \frac{p(t | pos)(1 - p(t | neg))}{p(t | neg)(1 - p(t | pos))} \end{aligned}$$

In the function, pos represent the positive instance and neg represents negative instances. ODD method does not treat all the class in the same as former evaluation functions, it focus the value of target class, which make ODD method is especially suitable for two-class classifier. In two-class task, it is hoped to distinguish the positive class but don't care about the negative class. But in real classification task, the negative instances usually occupy a percentage more than 90%, in this situation; the value of ODD has extra advantages than other information measures.

3. Experiment validation

3.1 Experiment Corpus

The publicly available benchmark corpora are used for our anti-spam filtering research in this paper, the 2005-Jun data set. The Chinese corpora we used is called 2005-Jun data set from China Education and Research Network Computer Emergency Response Team (CCERT), which is a non-profit organization who provides computer security related incident response service for people and organizations all over China. The 2005-Jun data set is the widest used corpora in China.

The 2005-Jun data set from CCERT contains 25088 spams and 9272 ham. A ham message is composed by a post (from forum) and a raw header of a ham messages. The raw data set removed all html tags in the

body part and kept only the plain text part of each message, but it remained the 'Content-type' header unchanged (it may be useful), but in our experiment, we removed the 'Content-type' header also. In another word, the Chinese corpora we used in our experiment are pure composed by Chinese characters.

3.2 Experiment Design

In order to compare the effect of the 6 algorithms, firstly, we preprocess corpus. Because of the inherent characteristics of Chinese corpus, we first delete the stop words and then create a global vocabulary, to save all the words appear on the corpus. We also use a hash table to record global DF, spam DF, and legitimate email DF. There is a part of terms in dictionary. Compare three methods of feature selection results; we design two experiments to compare the pros and cons of three methods on two different categorizing machines respectively. One set of experiments apply Bayesian methods for classification, the other SVM. And then do cross-validation. Then, we use six different feature selection methods, DF, IG and χ^2 statistic (CHI), Expected cross entropy (ECE), Weight of evidence for text (WET) and Odds ratio (ODD) to run feature selection for the documents. Some of the words selected are as shown. In addition to the selected words, the dictionary also records the score that word gains under different feature selection methods.

发票;0.778499	每月;0.158911
说;0.763479	商品;0.158607
很;0.59642	会;0.156173
合作;0.4888	普通;0.151195
都;0.438196	水;0.142512
开;0.420498	贵司;0.137442
优惠;0.358679	财务;0.134274
人;0.344316	左右;0.132113
代;0.337373	运输;0.126085
一个;0.298829	经理;0.125936
司;0.260758	电脑;0.12414
没;0.229652	信息;0.122805
联系;0.226304	社区;0.119874
服务;0.224995	现在;0.117903
想;0.209936	负责人;0.1159
没有;0.207602	进项;0.112588
销售;0.193705	增值税;0.111788
广告;0.191204	太;0.111505
贵公司;0.19007	一;0.110797
觉得;0.17359	站;0.109218
额度;0.168923	喜欢;0.100389

Figure 1. A part of terms chosen by IG

Figure 1. shows the terms which are chosen by IG algorithm. We increase the dimension from 20 to 200, in increments of 20. According to the different dimensions, we quantify of the spam, and finally generate VSM model, using WEKA experimental platform for classification.

3.3 Experiment evaluation

Four traditional measures in text classification research are used for evaluation: recall, precision, F1 measure and accuracy. For the following definitions let there be: a total of a messages, including a_1 spam and a_2 legitimate emails; b messages of which are judged as spam by an anti-spam filtering system, and among them, b_1 messages are really spam; and a total of a – b messages judged as legitimate, among which b_2 messages are really legitimate.

The evaluation measures of this system are defined as follows:

$$recall = \frac{b_1}{a_1}$$

$$precision = \frac{b_1}{b}$$

$$F1 = \frac{2 \times recall \times precision}{recall + precision}$$

F1 combines recall and precision into a single measure and is frequently used in TC. Accuracy is widely used in many previous anti-spam filtering studies. However, we believe that it is greatly affected by the class distribution. For example, if the number of legitimate messages is much bigger than that of spam messages, the accuracy will not be greatly affected by spam messages. In other words, the accuracy may be very high even though the spam filtering performance is very poor due to the skew distribution. So the F1 measure is mainly used in our experiments, but accuracy results are included for comparison with existing work.

3.4 Experiment result

3.4.1. Analysis of experiment result under NB platform: see Figure 2.

It can be seen from the figure that in accuracy comparison, the performance of ODD method is the most prominent, with the highest value reaching 0.958. This result is contrary to previous result in Newsgroup corpus using Bayesian classification methods. Previously IG performed better than DF. IG method

performs poorly, with the lowest value of 0.727. This result is also different with experiment result of multi-type problems we carried out before. In that experiment, we use Bayesian methods to compare IG in OHSUMED, showing IG performs better. This may be due to two reasons. First is the unbalanced nature of spam corpus, the other possibility is the difference of process in handling multi-type problems and two type problems. From this result we can see the ODD is good at dealing with two-class task in text classification.

Results from the recall rate show that ODD and χ^2 are obviously better than the other four methods. And the effect is more robust, with the highest value reaching 0.958.

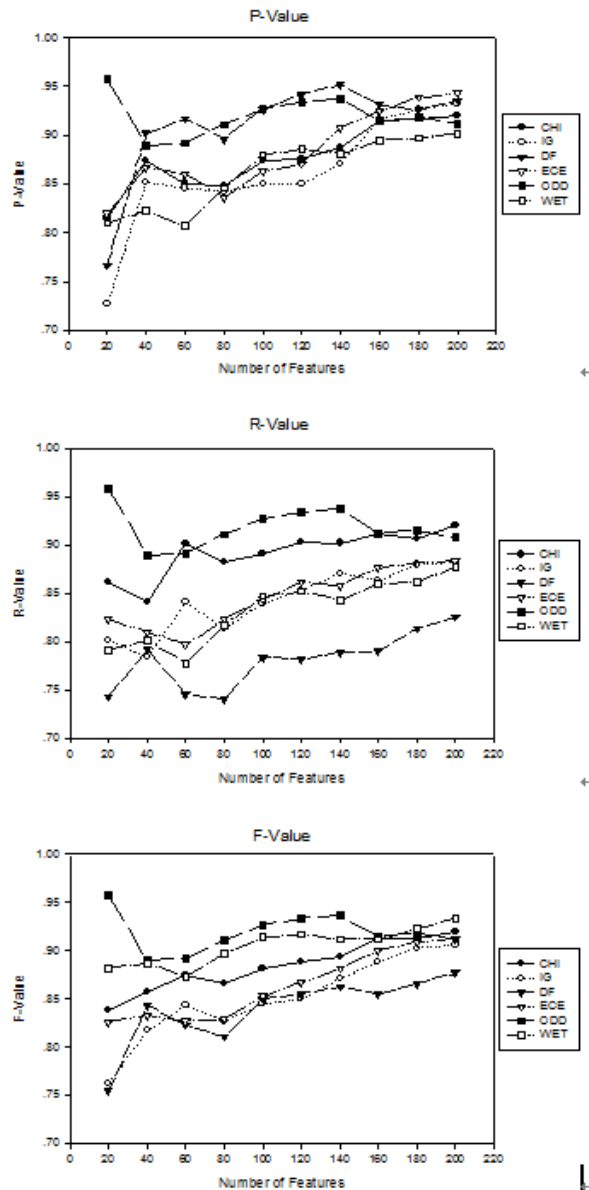


Figure 2. Results of 3 feature selection methods in Naïve Bayes

Comprehensive indicator of the P and R values of F1, we can see, in general, although the ODD method is not particularly advantageous, but is still better than the other methods.

3.4.2. Analysis of experiment result under SVM platform:

See Figure 3. the figure shows that in accuracy comparison, ODD and WET algorithms have more advantages. In Both high-dimension and low-dimension, it demonstrates its superiority, while the other four methods do not differ much. And we can also see that in the accuracy of comparison, the result is outstanding from our experiments, which can even

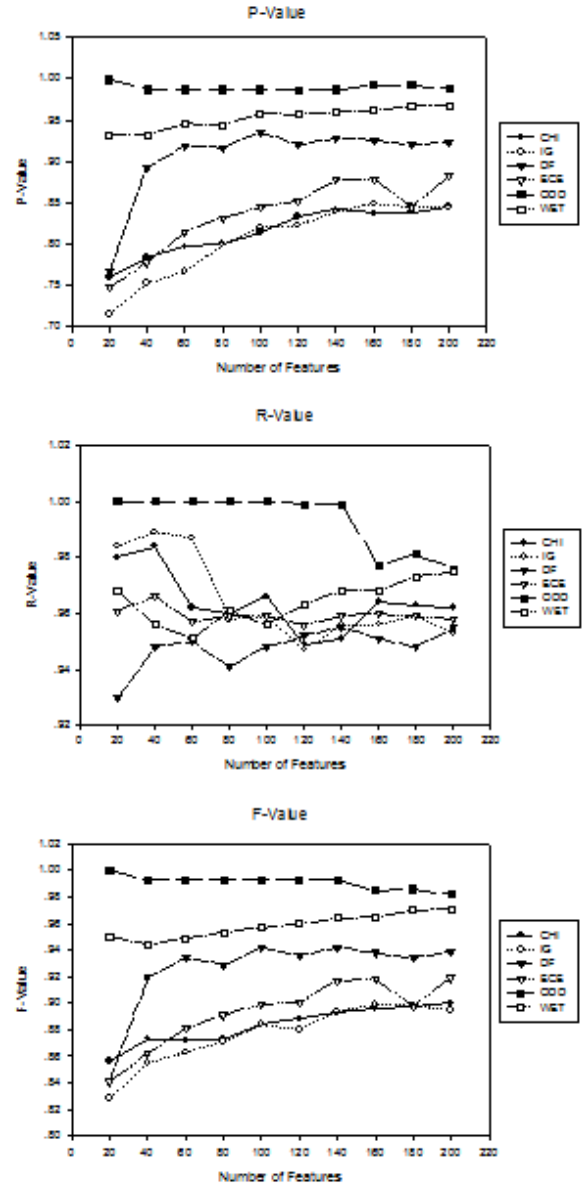


Figure 3. Results of 3 feature selection methods in SVM

reach 0.99 in precision. And it is also stable. Therefore we can conclude: ODD and WET has performed better than other feature selection methods on SVM platform. While the results of the recall rate have large ups and downs. ODD algorithm does not occupy too much advantage under high-dimensional space. Finally, when we look at the value of F1, ODD and WET still reflect a clear advantage in general, especially in the relatively low-dimensional spaces, while in higher-dimensional space, they are still better than the other four algorithm. This may be because that ODD and WET are better to deal this kind of data distribution while the χ^2 and IG are not as good, but also reflect good results.

4. Conclusion and future work

This is an evaluation of feature selection methods in dimensionality reduction for spam filtering at all the reduction levels of aggressiveness, we found ODD and WET most effective in aggressive term removal without losing categorization accuracy in our experiments with NaïveBayes and SVM.

The excellent performance of ODD, WET and indicates that these two feature selection methods are more suitable for binary classifier at least for our spam corpus. Because the characteristics of these two methods, they perform better than IG or CHI though they perform not so good in multi-class tasks. The availability of a simple but effective means for aggressive feature space reduction may significantly ease the application of more powerful and computationally intensive learning methods, such as neural networks, to very large spam filtering, which is otherwise intractable.

Our next task is to further analyze the performance of the good feature selection algorithm, find their commonalities, namely, feature selection algorithm in formal analysis text classification, and do more experiments on corpus with different language.

Acknowledgment

This paper is supported by Beijing Municipal Natural Science Foundation No: 4122076. And it is also one of the stage achievements of the study "Network based education of Chinese international promotion research" which is belong to Beijing education and science "the eleventh five years" planning, No: AHA091110.

References

[1]. Yiming Yang, Jan O. Pedersen. 1997. A Comparative Study on Feature Selection in Text Categorization. Proceedings of ICML-97, pp. 412-420.

[2]. Fabrizio Sebastiani Machine learning in automated text categorization [J]. ACM Computing Surveys, 34(1): 1-47. 2002.

[3]. Stewart M . Yang , Xiao-Bin W u , Zhi-Hong Deng, Ming Zhang, Dong-Qing Yang. 2002 Relative term-frequency based feature selection for text categorization [A] . Proceedings of ICMLC-2002[c]. 1432-1436.

[4]. Pawlak Z. Rough Sets. International Journal of Computer and Information Science, 1982, 11(5): 341-356

[5]. Xu Yan , Wang, Bin , Knowledge Measurement Based on Rough Set , 2009 IEEE INTERNATIONAL CONFERENCE ON GRANULAR COMPUTING , 2009/8

[6]. Andrew Moore. Statistical Data Mining Tutorials[DB/()I].<http://www.autonlab.org/tutorials/>

[7]. Yiming Yang, Xin Liu. A re-examination of text categorization methods [A.(SIGIR 99)[C].1999, 42-49.

[8]. H.Zhang. The optimality of naive Bayes [A].The 17th International FI AIRS conference [C] . Miami Beach: 2004. May 17-19.

[9]. Yiming Yang. An evaluation of statistical approaches to text categorization[J].Journal of Information Retrieval, 1 999, 1(1/2):67-88.

[10]. F. Cranor and B.A. LaMacchia. Spare! Communications of the ACM, 41(8):74--83, 1998.

[11]. Vaughan Nichols SJ (2003). Saving private e-mail. IEEE Spectr 40(8):40-44

[12]. China's anti-spam alliance <http://anti-spam.org.cn/> 2009

[13]. Symantec's report <http://www.symantec.com/business/theme.jsp?themeid=threatreport,2011>

[14]. Anti-Spam Committee is an affiliated organization of ISC <http://www.anti-spam.cn/> ,2009

[15]. Sahami M, Dumais S, Heckerman D, Horvitz E (1998) A Bayesian approach to filtering junk e-mail. In: Proceedings of AAAI workshop on learning for text categorization, pp 55-62

[16]. Androutsopoulos I, Koutsias J, Chandrinos KV, Spyropoulos CD (2000) An experimental comparison of Naive Bayesian and keyword-based anti-spam filtering with encrypted personal e-mail messages. In: Proceedings of the 23rd annual international ACM SIGIR conference on research and development in information retrieval, pp 160-167

[17]. Drucker H, Wu D, Vapnik VN (1999) Support vector machines for spam categorization. IEEE Trans Neural Netw 20(5):1048-1054

[18]. Carreras X, Marquez L (2001) Boosting trees for anti-spam email filtering. In: Proceedings of European conference on recent advances in NLP, pp 58-64

[19]. Andrej Bratko, Gordon V. Cormack, Spam Filtering Using Statistical Data Compression Models. Journal of Machine Learning Research 7 (2006) 2673-2698

Web Content Reauthoring for Large Screens

Ms. Neetu Narwal
Asst. Prof., Maharaja Surajmal Institute
Affiliate College of GGSIP University
New Delhi

Dr. Saba Hilal
Research Supervisor
HOD, Department of Computer Applications,
Lingaya's University,
Faridabad

Abstract—With advancement in the technology numerous display devices are available ranging from small screen i.e., palm top, androids, mobile phone etc., to the large screen devices such as 1200x1800, LCD display etc. In the small screen dimension lots of research had already been conducted but other end has received very little focus. We propose a methodology to adapt the web page on large screen context. We would be utilizing the native web technologies i.e., HTML5 and CSS3 to adapt the web page to large screen context and vision based methodology to extract the content of the web page

Index Terms—Web Page Segmentation, HTML5, CSS3.

I. INTRODUCTION

WITH the advancement in the technology in the last decade the use of Internet on the small scale device has gained momentum. There is a lot of research directed towards these category of devices i.e., small screen size, less bandwidth, limited processing power and restricted memory etc. However, recent years have also witnessed a drastic technological advancement in the large display devices and they have come within the reach of individuals. But the websites available on Internet do not utilize the large screen space available in these large screen devices.

The developers have produced several variations of the same web site to adapt to specific device category mainly desktops and smart phones. However, there is no remarkable advancement in the large display device category, still most of the websites do not utilize the available large screen space. Most of the websites are designed using fixed size position as the developer can't predefine the look of web site on different Web Browser so the use of fixed size positioning give them opportunity to provide consistent look to the web page across various browsers. This leads to wastage of space when these websites are accessed in large display devices.

II. PREVIOUS WORK

A. Research done in the area of Web Page Segmentation

Web data extraction system is a sequence of steps that extracts the content of the web page. There is lot of work

done in the area of extraction of content from web pages. There are several algorithms used for Web Page Segmentation, the most popular of which are

- DOM-based algorithms
- Layout-based algorithms
- Vision Based algorithms

DOM based web extraction method, make use of the Document Object Model [14] described by W3 Consortium. DOM Document is a collection of nodes arranged in a hierarchy, which allows a developer to navigate through the tree to extract the information.

Some approaches make used of HTML tags to divide the web page content [9,10] and some other rely on information like content [11] and link [12].

As this method is dependent on HTML tags and it has been observed that Web Designer does not follow very strict grammar rules for designing the web page so the DOM based method normally does not give accurate results.

The layout-based segmentation method uses layout information of the web page to relate the similar content blocks based on location, shapes and other parameters. Hiroyuki Sano et.al. [13] make use of layout template of a Web page for segmentation. Layout template is useful in terms of speculating where the main content of the page is located.

Cai et.al. [4, 5, 6] use layout information such as “font,” “color,” and “size” to restructure a Web page in a content block tree.

Cai, et.al. [5] have introduced a vision-based page segmentation (VIPS) algorithm. This algorithm segments a Web-page based on its visual characteristics of the web page, they segment the page based on colors, font, background etc. to identify the coherence of each blocks and identify the implicit and explicit separators to divide the web page into semantic blocks.

Nwe Nwe Hlaing, et.al., [8] used VIPS algorithm to extract the blocks from the web page and then analysed the data regions to extract the data records proposed Vision based Extraction of data Record (VER) algorithm.

B. Some of the research has been done in the area of Web Page adaptation for Large Screen Context

Michael Nebeling et.al.[1], have developed a set of metrics that can be used understand the utilization of space in different widescreen settings. These metrics can be used by developer to design a web page for large and high resolution screens.

Michael Nebeling[2] et.al, suggested an adaptive layout template that can accommodate a range of viewing situations and because it is based on only native web technologies, can be easily applied to existing web sites. They made use of HTML5 and CSS3 specification such as advanced media queries and multi-column layout for adaptive viewing.

Nobuo et.al.[3], suggested an extension of Web-page layout optimization method for multimodal browsing sizes, it dynamically changes box locations and font sizes by switching CSS files for different browsing sizes, so that the main content can be accessed without the screen scroll operation even at a small sizes whereas the blank space is avoided at a large size.

III. PROPOSED METHODOLOGY

We propose a methodology to adapt the web page according to the large screen display by dynamically repositioning of the elements from the web page to make use of available space. This methodology would avoid redesigning of the websites for different web browsing devices. The plugins can be attached with the web browser for rendering the webpage blocks to adapt themselves according to screen context.

We propose the following steps in adapting the web page and re-authoring for large screen devices context.

Step 1: The first step would be Web Data Extraction, according to literature survey these are broadly categorized as DOM based, Layout based and Vision based. We are using vision based method for our work.

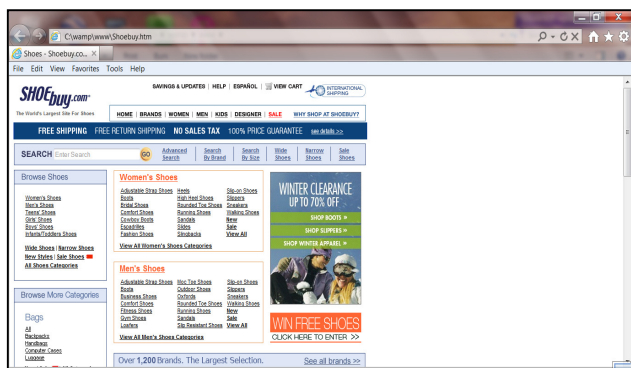


Fig 1 : Original Web site displayed in the Web Browser on large screen display

Visual based method is based on visual cues of the web site i.e, color, font, background etc. and also make use of DOM tree for segmentation of the web page. It generates visual block tree as the output.

Vision based method may incorporate three aspects for measuring the distance of elements for grouping them in blocks

- i) Geometric measure
- ii) DOM measure
- iii) Semantic measure

Our proposed method makes use of Vision based method for web page segmentation and using all three measures for calculating the degree of coherence between blocks and then dividing them into visual blocks.

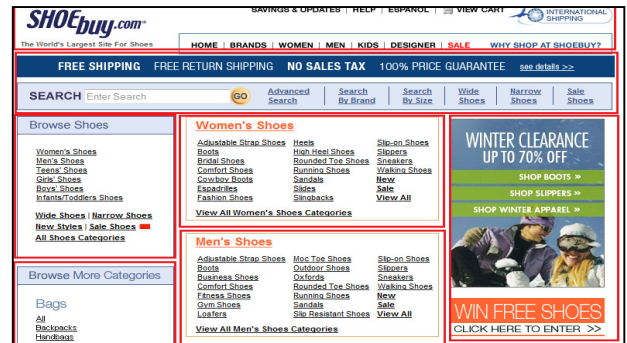


Fig 2 : Visual Blocks Captured from the Web Site

Step 2: The next step is to redesign the style sheet so as to accommodate the web page according to the window viewport size and the window size can be obtained by using CSS3 media queries. We will be making use of CSS3 and HTML5 property i.e, multi-column layout, float property, liquid layout etc. for rearranging the blocks of element to utilize the available extra screen space.

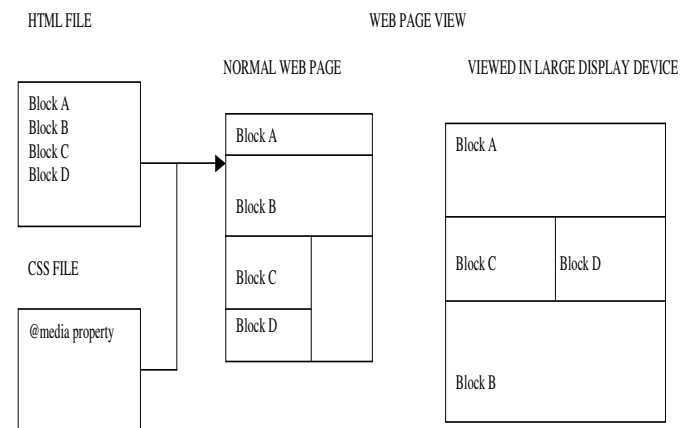


Fig3. Block Ordering and display format changed by accordingly in large display device

New style rules will be added to already existing style sheet if no style sheet is found then new .css file will be

created and linked to the current web page, the rules will be generated by keeping in mind the semantic similarity of element blocks.

The media style rules, normal flow, the liquid layout and the float property of CSS will be used for adjusting the blocks according to the size of display device

Step 3: After adding the style rules and rearranging the web page element blocks, the web page will be reloaded utilizing the extra screen space of large display device.

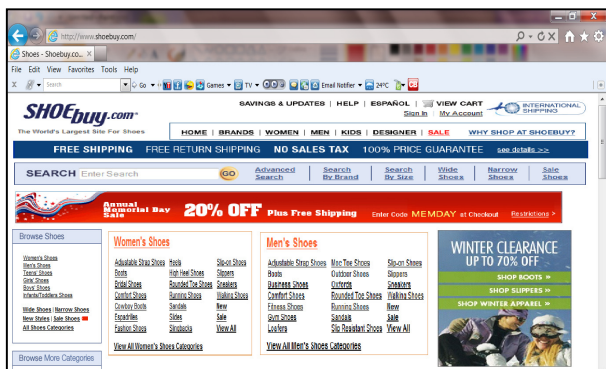


Fig 4 : Web Site after rearranging the visual blocks

CONCLUSION

This paper presents a proposed scheme for Web page adaptation for large screen context. The method makes use of vision based techniques for web data extraction and also make use of media queries of CSS3 for modifying the web page content based on the window viewport size. The result has not been presented in this paper as the work is still in progress.

REFERENCES

- [1] Michael Nebeling, Fabrice Matulic, Moira C. Norrie , Metrics for the Evaluation of News Site Content Layout in Large-Screen Contexts, , Vancouver, BC, Canada, 2011.
- [2] Michael Nebeling, Fabrice Matulic, Lucas Streit, and Moira C. Norrie Adaptive Layout Template for Effective Web Content Presentation in Large-Screen Contexts,
- [3] Nobuo Funabiki, Junki Shimizu, Megumi Isogai, Toru Nakanishi , An Extension of the Web-page Layout Optimization Method for Multimodal Browsing Sizes, 13th International Conference on Network-Based Information Systems, 2010
- [4] Deng Cai, Shipeng Yu, Ji-Rong Wen and Wei-Ying Ma, Extracting Content Structure for Web Pages based on Visual Representation, , Microsoft Research Asia
- [5] Deng Cai, Shipeng Yu, Ji-Rong Wen and Wei-Ying Ma, VIPS: A Vision based Page Segmentation Algorithm, , Microsoft Research Asia, Beijing, China, 2004
- [6] Deng Cai, Xiaofei He, Ji-Rong Wen, Wei-Ying Ma, Block-level Link Analysis, , ACM 2004
- [7] Suhit Gupta, Gail Kaiser, David Neistadt, Peter Grimm, DOM-based Content Extraction of HTML Documents, Budapest, Hungary, 2003.
- [8] Nwe Nwe Hlaing, Thi Thi Soe Nyunt, An Approach for Extraction Data Record from Web Page based on Visual Features, IJAMS, Aug 2011.
- [9] Lin, S.-H. and Ho, J.-M., Discovering Informative Content Blocks from Web Documents, In Proceedings of ACM SIGKDD'02, 2002.

- [10] Wilkinson, R., Effective Retrieval of Structured Documents, In Proceedings of the Seventeenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Dublin, 1994, pp. 311-317.
- [11] Embley, D. W., Jiang, Y., and Ng, Y.-K., Record-boundary discovery in Web documents, In Proceedings of the 1999 ACM SIGMOD international conference on Management of data, Philadelphia PA, 1999, pp. 467-478
- [12] Chakrabarti, S., Joshi, M., and Tawde, V., Enhanced topic distillation using text, markup tags, and hyperlinks, ACM Press, 2001, pp. 208-216.
- [13] Hiroyuki Sano, Shun Shiramatsu, Tadachika Ozono, and Toramatsu Shintani, A Web Page Segmentation Method based on Page Layouts and Title Blocks
- [14] Web Site Citation : <http://www.w3.org/DOM/>

Feature Based Iris Recognition System Functioning on Extraction of 2D Features

Arjun Agrawal
Systems Engineer
Infosys Limited
Bangalore, India
arjun.4060@gmail.com

Gundeep Singh Bindra
Department of Computer Science
SRM University
New Delhi, India
mailbox@gundeepbindra.com

Priyanka Sharma
Department of Computer Science
Global Institute of Technology
Jaipur, India
Priyanka.sharmagit@gmail.com

Abstract—In this paper a novel and simple iris feature extraction technique is proposed for iris recognition of high performance. We use one dimensional circular ring to represent iris features. The reduced and significant features afterward are extracted by Sobel Operator and 1-D wavelet transform. To improve the accuracy, this paper combines uses Euclidean Distance for classification.

Keywords—: *iris recognition; wavelet transform; probabilistic neural network; particle swarm optimization*

1. INTRODUCTION

Among all the biometric indicators, iris has one of highest levels of reliability. The human iris, an annular part between the pupil and the white sclera as shown Fig. 1(a), has an extraordinary structure and provides many interlacing minute characteristics such as freckles, coronas, stripes, etc. These visible characteristics, which are generally called the texture of the iris, are unique to individuals. Even identical twins having similar DNA, are believed to have different iris.

Iris recognition [1-13] is the process of automatically differentiating people on the basis of individuality information from their iris images. The technique can be used to verify the identity of a person when accessing a system. Due to its reliability and high precision, it is beneficial for biometric authentication system.

A typical iris recognition system can be composed into three modules: an iris detector for detection and location of iris image, a feature extractor and a matcher module.

In this paper, we focus our investigation on a new iris detector; feature extraction and representation approach to further implement an iris recognition system with low complexity and high performance.

Firstly, in order to reduce system complexity, we

use 2-D wavelet transform [14] to obtain a low resolution image and localize pupil position. By the center of pupil and the radius of pupil, we can acquire the iris circular rings. The more iris circular rings are acquired, the more information is abundant. Secondly, we segment the iris image into three parts and two parts. In each segmented iris image, iris texture is extracted as feature vector by Sobel operator. The 1-D discrete wavelet transform is adopted to reduce the dimensionality of the feature vector. In our experiments, the wavelet permits to further reduce the system complexity and obtain a discriminated feature vector. We use Euclidean Distance approach for classification problems. Finally, the combination of the novel feature extraction method and Euclidean distance classifier is evaluated on the IITK iris database for iris recognition. The experiment results present that the proposed method is well suitable for a low complex computation and low power devices.

2. DETECTION OF IRIS REGION

The iris image, as shown in Fig. 1 (a), contains not only abundant texture information, but also some useless parts, such as eyelid, pupil, etc. The iris is between the pupil (inner boundary) and the sclera (outer boundary). In order to locate the pupil, a simple and efficient method is proposed. The procedure is as following:

1. Take a raw image and apply 2-D wavelet filtering. The size of the resulting image is only quarter of the original image.
2. Compute the histogram to find the maximum peak.

The Fig. 1(b) shows that the maximum peak of histogram is the gray values of pupil region, because the pupil region is concentrated on the lower gray values. The maximum peak is set to P and the threshold T is obtained by P×W. The W is weight and the value of it is set to 1.1 in the paper. The binary image B is obtained by the original image A.

$$B(i, j) = \begin{cases} 1, & \text{if } A(i, j) > T \\ 0, & \text{if } A(i, j) < T \end{cases}$$

Because the binary image B has still some black points outside the pupil region, an estimated point is computed by a function E(i, j).

$$E(i, j) = \prod_{ii=-1}^1 \prod_{jj=-1}^1 B(i+ii, j+jj)$$

And

$$B'(i, j) = \begin{cases} B(i, j), & \text{if } E(i, j) > 4 \\ 1, & \text{otherwise} \end{cases}$$

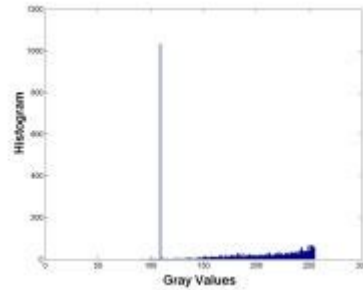
If the value of E(i, j) is greater or equal to 4, there are at least four dark points surrounding the B(i, j). Because the estimated point is located in pupil region, the estimated point is retained. Otherwise the estimated point is removed. Finally, we use the vertical projection and horizontal projection to obtain the center coordinates and the radius of pupil. The extracted pupil region is shown in Fig. 1(c).

3. Because the center coordinates and the radius of the pupil are multiplied by two, the center coordinates and the radius of the pupil are obtained in original eye image.
4. The iris circular ring (such as seen in Fig. 2 (a)) is obtained by giving a radius from the center of the pupil.

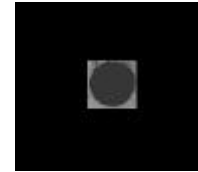
In the above mentioned procedure, the first step reduces the dimensionality of image to improve the efficiency of iris image extraction. The second and third steps provide an approach to locate the position of the pupil.



(a)



(b)



(c)

Fig 1.(a) Original image (b) The histogram of the gray values (c) Segmented pupil

We give different radius to get iris circular rings of different size. The more iris circular rings are extracted, the more information is used as features. The recognition performance is much better, but the efficiency is slightly affected. The proposed method is different from the traditional methods. The traditional methods extract a complete iris image, but the proposed method only extracts parts of the iris image for recognition. This will result in lower computational demands. In the next section, the detailed description of the iris feature extraction method will be presented.

3. IRIS TEXTURE FEATURE EXTRACTION

We extract consecutive circular rings using step 4 of iris location procedure. The more iris circular rings are extracted, the more information is used as features. The recognition performance is much better, but the efficiency is slightly affected. The proposed method is different from the traditional methods. The traditional methods extract a complete iris image, but the proposed method only extract parts of the iris image for recognition. This will result in lower computational demands.

A) Circular-derived iris blocking image

These circular rings then are stretched horizontally and accumulated, and construct a rectangular-type iris block image, shown as in Fig. 2 (b). Iris texture has abundant texture information for iris recognition. Here we elaborate a very simple and fast algorithm to extract iris feature for iris recognition. We previously proposed a novel iris feature extraction [13], but the recognition performance is not good.

In order to improve the recognition performance, the iris image is divided into three parts (see Fig. 2 (c)) and two parts (see Fig. 2 (d)). In order to compensate for a variety of lighting conditions, the segmented iris image is normalized (see Fig. 2 (e)).

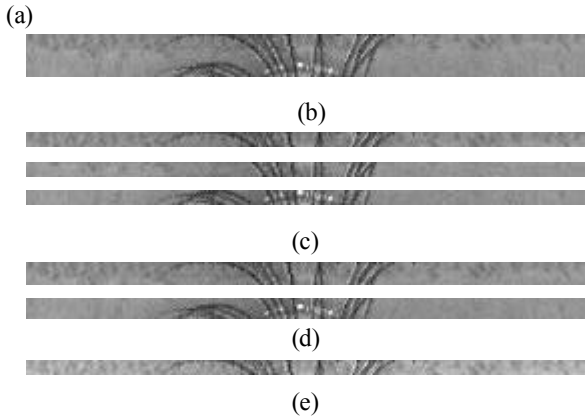
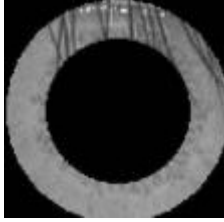


Fig.2 (a) original iris image (b) stretched iris block image;(c) iris image divided into three parts; (d) iris image divided into two parts; (e) normalized iris image

B) Sobel operator

The iris image is captured in different size from different people. It is not convenient for iris recognition, and the recognition performance is also affected. In the cause of the convenience of computation and achieving the high recognition performance, the number of captured iris circular ring from different iris image is the same. In order to enhance the texture of iris, the iris image is normalized. We adopt the Sobel operator to analyze texture shown as in Fig. 3 and the Sobel mask S_x is as following:

$$S_x = \begin{bmatrix} -1 & 0 & 1 \\ -2 & 0 & 2 \\ -1 & 0 & 1 \end{bmatrix}$$



Fig 3. Iris image after Sobel operator

C) Vertical projection

We adopt vertical projection to obtain 1-D energy profile signal and to reduce system complexity. In order to concentrate the energy, every row is accumulated as energy signal.

Let G be a segmented iris image of size $m \times n$, m is the number of iris circular ring, and n is pixels of each iris circular ring.

$$G = \begin{bmatrix} g_{1xn} & \dots & g_{1xn} \\ \vdots & \ddots & \vdots \\ g_{1xn} & \dots & g_{1xn} \end{bmatrix}$$

After vertical projection, the 1-D energy signal S is obtained.

$$S = [s_1 \ \dots \ s_n]$$

The m is much smaller than the n . Thus, the information of iris texture after vertical projection is more than the information after horizontal projection. So, we adopt the vertical projection to extract the 1-D energy signal.

D) Wavelet transform

The wavelets [14] to signal and image processing have provided a very flexible tool for engineers to apply in various fields such as speech and image processing. In an iris recognition system, the 2-D wavelet transform is only used for preprocessing. The preprocessing helps to reduce the dimensionality of feature vector and to remove noise. Nevertheless, the computational complexity is comparatively high. Thus, the paper proposes 1-D wavelet transform as filter to reduce the dimensionality of feature vector, and it can further reduce the computational complexity.

The wavelet is constructed from two-channels filter bank as shown in Fig. 4 (a). In wavelet decomposition of 1-D signal, a signal is put through both a low-pass filter L and a high-pass filter H and the results are both low frequency components $A[n]$ and high frequency components $D[n]$. The signal $y[n]$ is reconstructed by the construction filters H and L .

The wavelet filters are used to decompose signal s into high and low frequency by convolution.

$$D[n] = \sum_{k=-\infty}^{\infty} s[k] \cdot H[n-k] \quad D = s, H$$

$$A[n] = \sum_{k=-\infty}^{\infty} s[k] \cdot L[n-k] \quad A = s, L$$

In order to construct multi-channel filter, we can cascade channel filter banks. Fig. 4 (b) represents a 3-level symmetric octave structure filter bank. This is an important concept from multi-resolution analysis (MRA).

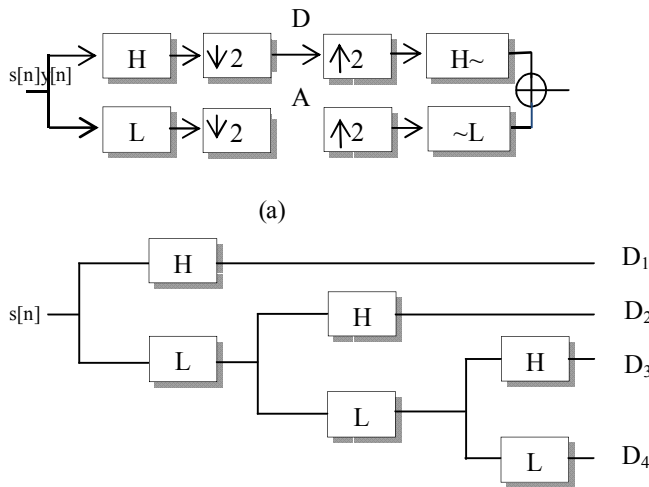


Fig.4. (a) Two-channels filter bank (b) 3-level octave band filter bank

4. FEATURE MATCHING

Feature Matching corresponds to study how iris can be identified. The matching between input image and image already Stored in database are matched. The feature vectors are matched in two images to conclude about the similarity and dissimilarity of the image.

The technique used for matching feature vectors is the Euclidean distance (Euclidean distance is the “ordinary distance” between two points that one would measure with a ruler and is given by the Pythagorean formula) based approach, unlike the traditional probabilistic neural network (PNN) approach using PSO, a new bio-inspired optimization method, as the learning algorithm. We calculate the Euclidean distance between the feature vectors and smaller is the distance, less is the dissimilarity between the two vectors.

Using various values of threshold, we calculate False Acceptance Rate (FAR) and False Rejection Rate (FRR) and hence get the Equal Error Rate (EER), as the intersection point of FAR and FRR curve. The high performance of iris verification system is in low Earthen testing images are matched with training set images. Using various values of threshold, we calculate False Acceptance Rate (FAR) and False Rejection Rate (FRR) and hence get the Equal Error Rate.

5. EXPERIMENTS & RESULTS

The entire algorithm was implemented using MATLAB 7.6.0.324(R2008a) .A set of iris images taken from 102

individuals was worked upon and tested. The training and testing of images was done. In order to understand the implementation of the algorithm we performed segmentation of iris images .Presently we have worked upon the segmentation in two and three parts. The study was done by using sobel operator and without using sobel operator. The results are as follows :

- 1) Results for two segments Image using sobel operator

Accuracy = 9.090909e+001=90.90909%
Equal Error Rate = 1.869345e+001=18.69345%
 FRR @ 1/100*FAR = 1.421195e-002
 FRR @ 1/1000*FAR= 4.737316e-003
 FAR @ 1/1000*FRR= 0

- 2) Results for two segments Image without sobel operator

Accuracy = 9.043062e+001= 90.43062%
Equal Error Rate = 2.011938e+001=20.11938%
 FRR @ 1/100*FAR = 1.894926e-002
 FRR @ 1/1000*FAR= 9.474632e-003
 FAR @ 1/100*FRR= 0
 FAR @ 1/1000*FRR= 0

- 3) Results for three segments Image using sobel operator

Accuracy =9.282297e+001=92.82297%
Equal Error Rate = 1.676536e+001=16.76536%
 FRR @ 1/100*FAR = 1.421195e-002
 FRR @ 1/1000*FAR= 4.737316e-003
 FAR @ 1/100*FRR= 0
 FAR @ 1/1000*FRR= 0

- 4) Results for three segments image without using sobel operator

Accuracy = 8.755981e+001=87.55981%
Equal Error Rate = 2.060732e+001=20.60732%
 FRR @ 1/100*FAR = 1.894926e-002
 FRR @ 1/1000*FAR= 9.474632e-003
 FAR @ 1/100*FRR= 0
 FAR @ 1/1000*FRR= 0

REFERENCES

- [1] J. Daugman, *Biometric Personal Identification System Based on Iris Analysis*, United States Patent, no. 5291560, 1994
- [2] J. Daugman, "High Confidence Visual Recognition of Person by a Test of Statistical Independence," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 15, no. 11, pp1148-1151, Nov. 1993
- [3] C. Sidney Burrus, Ramesh A. Gopinath, Hai Ta Guo, "Introduction to Wavelets and Wavelet Transforms: A Primer", 1998.
- [4] J. Daugman, "Demodulation by Complex-Valued Wavelets for Stochastic Pattern Recognition," *Int'l J. Wavelets, Multiresolution and Information Processing*, vol. 1, no. 1, pp. 1-17, 2003.
- [5] J. Daugman, "Statistical Richness of Visual Phase Information: Update on Recognition Persons by Iris Patterns," *Int'l J. Computer Vision*, vol. 45, no. 1, pp. 25-38 2001.
- [6] Li Ma, Tieniu Tan, Yunhong Wang, and Dexin Zhang, "Personal Identification Based on Iris Texture Analysis", *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 25, no. 12, pp. 1519-1532, Dec. 2003
- [7] R. Wildes, J. Asmuth, G. Green, S. Hsu, R. Kolczynski, J. Matey, and S. McBride, "A Machine-Vision System for Iris Recognition," *Machine Vision and Application*, vol. 9, pp. 1-8, 1996.
- [8] R. Wildes, "Iris Recognition: An Emerging Biometric Technology," *Proc. IEEE*, vol. 85, pp. 1348-1363, 1997.
- [9] Christel-Loic Tisse, Lionel Torres, Michel Robert, "Person Identification Based on Iris Patterns," *Proceedings of the 15th International Conference on Vision Interface*, 2002.
- [10] L. Ma, Tan, Tieniu, Wang, Yunhong, "Iris Recognition Using Circular Symmetric Filters", *Processing of THE 16th International Conference on Pattern Recognition*, vol.2, pp. 414-417, 2002.
- [11] Mayank Vatsa, Richa Singh, and P. Gupta, "Comparison of Iris Recognition Algorithms", *Proceedings of ICISIP'04, India*, pp.354-358, 2004.
- [12] Sanchez-Avila C., Sanchez-Reillo R.; de Martin-Roche D., "Iris Recognition for Biometric Identification Using Dyadic Wavelet Transform Zero-Crossing", *Proceedings of the IEEE 35th International Carnahan Conference on Security Technology*, pp. 272-277, 2002.
- [13] Ching-Han CHEN, Chia-Te CHU, "High Efficiency Iris Feature Extraction Based on 1-D Wavelet Transform", *2005 Design automation and test in Europe (DATE2005)*, Munich, Germany, March, 2005.
- [14] Jaideva C. Goswami and Andrew K. Chan "Fundamentals of Wavelets" 1999.
- [15] D.F. Specht, "Probabilistic Neural Network for Classification, Map, or Associative Memory", *Proceeding of the IEEE International Conference on Neural Network*, vol.1, pp525-532, 1988.
- [16] Chia-Te CHU and Ching-Han CHEN, "The Application of Face Authentication System for Internet Security Using Object-Oriented Technology", *Journal of Internet Technology*. (Accepted)
- [17] Ching-Han CHEN, Chia-Te CHU, "An High Efficiency Feature Extraction Based on Wavelet Transform for Speaker 2004 International Computer Symposium (ICS2004)", Taipei, Dec 2004.
- [18] Ching-Han CHEN, Chia-Te CHU, "Combining Multiple Features for High Performance Face Recognition System", *2004 International Computer Symposium (ICS2004)*, Taipei, Dec 2004.
- [19] Ching-Han CHEN and Chia-Te CHU, "Real-Time Face Recognition Using Wavelet Probabilistic Neural Network", *Journal of Imaging Science and Technology*. (Accepted)
- [20] J. Kennedy et al. "Particle Swarm Optimization", *Proc of IEEE Int. Conf. Neural Networks*, vol. IV, pp.1942-1948, 1995.
- [21] CASIA iris database. Institute of Automation, Chinese Academy of Sciences, [Online]. <http://www.sinobiometrics.com/casiairis.htm>.

The Bus Arrival Time Service Based on Dynamic Traffic Information

Tongyu Zhu, Jian Dong, Jian Huang, Songsong Pang, BoWen Du
State Key Laboratory of Software Development Environment

Beihang University

Beijing, China

{ zhutongyu, dongjian, huangjian, pangsongsong, dubowen }@nlsde.buaa.edu.cn

Abstract—The bus arrival time (BAT) service is a key service to improve public transport attractiveness by providing users with real-time bus arrival information which can help them to arrange their bus travel schedule intelligently. Thus the technique of real-time bus arrival prediction has become a research hotspot in the community of Intelligent Transport Systems (ITS) nowadays. In this paper, a novel model on bus arrival time prediction is proposed. The model proposes a complete set of programs to solve BAT prediction for large-scale real-time traffic information calculating. It adopts an effective algorithm judging bus's driving direction real-timely. BAT is calculated based on dynamic traffic information and visual prediction is a way to complement when GPS information is not arrived as expected. Experimental results indicate that the model has considerable efficiency in accuracy (over 85.1%) and computational speed (max 5000 GPS records per second) .

Keywords—bus arrival time prediction; GPS data; dynamic traffic information;

I. INTRODUCTION

Traffic congestion has been increasing worldwide as a result of increased motorization, urbanization, population growth, and changes in population density. One way of addressing this problem is by providing more infrastructure to meet the increasing number of vehicles. However, there is a limit to this solution and hence other alternative options need to be explored to meet the growing traffic demand. One such option is better management of the existing facilities using Intelligent Transportation System(ITS).

Advanced Public Transportation System (APTS) is one of the most important ITS applications which can help in relieving congestion by attracting more people to public transport by improving the efficiency of public transportation systems. One such APTS application will be to provide accurate information about bus arrival to passengers, leading to reduced waiting times at bus stops. This requires the prediction of travel time, which is the total elapsed time of travel, including bus stop and intersection delay, necessary for a vehicle to travel from one point to another over a specified route under existing traffic conditions.

Historically, many researchers have adopted various methods such as historic and real-time approaches, machine learning techniques (Artificial Neural Network (ANN), support vector machines), model based approaches (Kalman filtering) and statistical methods (regression analysis, time-series) for the

prediction of bus arrival time. Lin and Zeng developed an algorithm to provide real-time bus arrival information using the historical approach [1]. They used the bus location data, the schedule information, the difference between scheduled and actual arrival times, and the waiting time at time-check stops. Wall and Dailey proposed an algorithm which used a combination of Automatic Vehicle Location (AVL) and historical data for bus arrival prediction [2]. The algorithm had two components: tracking (using Kalman filter) and prediction (using statistical estimation). Chien et al. developed an ANN model to predict dynamic bus arrival time [3]. They used simulated data from CORSIM including volume and passenger demand. Jeong and Rilett evaluated the performance of historical data based model, regression model and ANN model for bus arrival time prediction and reported the ANN model performing better than the other two models [4]. Cathey and Dailey prescribed an algorithm for bus arrival time prediction which had three components namely tracker, filter and predictor [5]. The Kalman filter was used in the filter component. Liu et al. developed a hybrid model based on State Space Neural Networks (SSNN) and the Extended Kalman filter (EKF) [6]. The SSNN model requires large data set for offline training. They developed an EKF model to train the SSNN.

The present study provides a model based on GPS information and algorithm for the bus travel time taking into account travel time between stations, dwell times at stops and delays at intersections. In order to ensure the high accuracy while keeping the efficiency of the basic data (such as bus line, station data, up and down direction) pretreatment, this paper proposes a complete set of solution to find bus line automatically, ensure station location, correct up and down direction of the bus real-timely. To meet the real time performance of the calculation for the bus GPS, a new algorithm searching projecting matched link is adopted. Furthermore, virtual prediction is adopted in the model when real-time GPS is not arrived as expected. All of this compose a complete model predicting bus arrival time and distance.

II. DATA PREPROCESSING AND EXTRACTION

The BAT prediction system processes GPS information of 18000 buses in Beijing real-timely. In consideration of the differences of road information between the suburb and urban, route 944 was chosen as the research route for both covering enough lines in suburb and urban. It is one of the most frequently used routes in Beijing and consists of 35 buses stops

and 15 signalized intersections with a travel time of approximately 80 minutes. Travel time related data is collected using handheld/onboard GPS instruments in all 944 line buses. The GPS Online Rate (GOR) reached 85%.

A. the calculation and correction of the line and station data

In this paper, the geodetic coordinates of link chain endpoint was deposited in route files in turn according to the order of a bus pass, which provided a condition for bus accurate positioning and calculation of real-time distance to station. Meanwhile each link was ordered uniqueness by bus passing in order to distinguish repeatedly passing the same link of line.

Link is described as follows:

$$L = \{l = (id, length, sCoor, eCoor, samplingpoints)\} \quad (1)$$

There is always a key id that indicates a directed link l , $startCoordinate$ is the start point of the link l , $endCoordinate$ is the end point of the link l . $samplingpoints$ are a set of GPS which include longitude and latitude.

Bus line is described as follows:

$$BL = BL(B, D, L) \quad (2)$$

L is a set of Links ordered exclusively by the sequence of link bus passed by. Given $\forall b \in B, d \in D, l \in L$ a bus line $o(B, D, L)$ can be obtained from BL .

On account of enormous amount of the number of lines, a lot of manual work will be needed if all the work is completed in manual. Clustering of the geodetic coordinates for link bus pass by makes the bus line link set founded automatically for every single bus line. Through clustering of the geodetic coordinates of GPS information of running bus, the program will find road link automatically and store them into bus line file. Secondly, manual proofreading ensures the correctness of the results.

Meanwhile, the station location data acquired from official agency are not consistent with actual station location. The deviation will lead to 100 meter difference for bus arrival distance. In order to find the accurate station location, three ways were proposed to find and fix station location result.

First, Station Location Information (SLI) can be got from the official database. And then SLI can also be obtained from google map. At last, due to bus always stop or slow down at station, where bus travels in a low speed, the number of bus GPS point around the station is high and intensive.

Based on above analysis, spatial clustering of the bus GPS point is applied to located accurate bus station after wiping out exception data point. Through mutual verification of the methods and results contrasts, station location can be identified.

B. Driving Direction and Next station Calculation Algorithm

All buses in Beijing have been equipped with GPS instruments. The real-time positioning information of bus is mainly got from onboard GPS instruments. The Real-time

GPS information for one bus arrived in every 20 seconds called a transfer period T , which contains the company name, License plate number, longitude, latitude.

Vehicle Number, downstream and upstream, longitude, latitude and next station number are basic data sources for calculating bus arrival distance and time. The quality of DAU, next station number, longitude and latitude have a significant impact on the accuracy of the calculation result. Considering BAT services system's strong dependency on above data, an algorithm needs to be proposed to calculate basic data and guarantee the accuracy.

First, GPS offset data need to be projected to the link by vertical matching. The pedal is the matched result.

Second, line data and bus station data need to be processed first before calculating the DAU and next station stop. The line corresponding to the sampling points are numbered. Then 5 real-time GPS points are coached in program and store the new data and matched it to the link. Bus station data after handling contains station name, station number, DAU, matched link, difference, matched sampling point, station longitude, latitude.

Sampling point is described as follows:

$$s = (longitude, latitude) \quad (3)$$

The matched sampling point number can be described as:

$$v(b, t, sn) \quad (4)$$

For a given bus b and specified timestamp t , $v(b, t, sn)$ is the matched sampling point number. For every bus, a vector of $v(b, t, sn)$ is coached which hold 5 points.

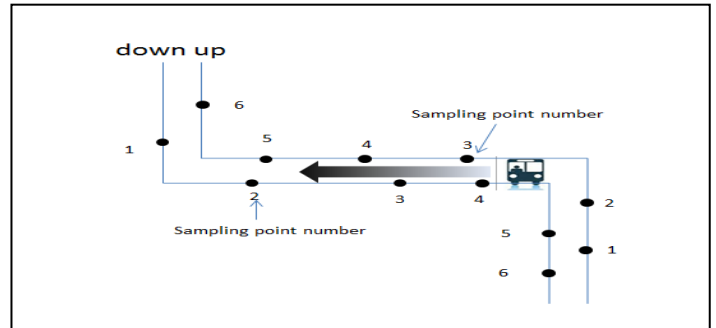


Figure 1. Model used to judge up and down direction

For a certain bus, the direction of bus running is judged by the order of 5 coached points sorted by timestamp. Assuming t_5 means the newest coached point, t_4 is last to t_5 . The direction bus is traveling is upstream if the order of the $v(b, t_5, sn)$ and $v(b, t_4, sn)$ is same with upstream. In reverse, the direction is downstream. As shown in Fig.1, the order is $\{1, 2, 3\}$ which is same with the sampling number sequence of upstream the bus has passed by. Through judging the number sequence of the bus station and the coached GPS point, the station number is calculated the current bus is running to.

However, the algorithm discussed above takes 3 seconds to calculate the result in the actual testing process. Given 20s

interval of the data dissemination , 15 seconds is devoted to other process which lead to 5 seconds left to handle 9000 data 10000 buses. Therefore the algorithm can not support real-time data calculation under current data scale.

The time of the algorithm mainly spend on matching the nearest sampling point. Getting the matched result needs calculating distance between the bus GPS point and all sampling GPS points by turn. Assuming that a line consists of m sampling points, the time complexity of the algorithm is $O(n*m)$ for n data one time matched.

For matching the nearest point faster, sampling points are stored in a two-dimensional array firstly. Then, data in the two-dimensional array is rearranged by. The ranked time complexity is $m \log m$. For certain x-coordinate and a given distance 100 m threshold which the nearest distance between the bus GPS point and sampling point is less than, binary chop is executing on x-coordinate for finding sampling point x_f in the array with the smallest difference between x-coordinate in the bus GPS point and x-coordinate in sampling point. Based on the point x_f found above, the algorithm give a search in a extending the range $(x_f.x - 0.0001, x_f.x + 0.0001)$ (100m extends the scope from the x_f in horizontal ordinate). The algorithm gives a 25 times efficiency improved and 100 times efficiency improved especially for east-west line. 1 second is taken handling 10000 data after applying the new algorithm.

C. GPS and Dynamic traffic information

The distance traveled in each of transfer period T is calculated from the corresponding latitude and longitude using the Haversine formula. Thus, the processed data included the distance and the corresponding travel time between consecutive locations for all buses and the delay values which were noted down manually.

This research utilizes Beijing's Dynamic Traffic Information Service System, constructed by CENNA VI Corporation. This system obtains traffic information through over 140,000 float cars and 5,000 buses and publishes the information every 2 minutes. The published information includes: congestion, travel time, state and emergency for each link. This research also utilizes Beijing's history traffic information, which publishes every 2 minutes and updates every week.

III. DEVELOPMENT OF PREDICTION MODEL

The bus arrival time prediction is a very complex problem, which involves many factors influenced by random. The model improves the accuracy of prediction model at the same time contents and creates various need for the customer. The prediction model of bus arrival time is divided into two categories. The first category is short forecast. The model will give a forecast for the next 3 stations the bus will be passed by. The second category is long-distance forecast which predicts bus arrival time for the stations 3 stations away from the current location. The two models will be introduced respectively as follows:

A. Short distance prediction

When the bus is within three bus stations from the current site, the approach presented here is to divide the total travel time of a bus into three components - its running time, its dwell time at bus stops and its signalized intersection delay time. The entire section of travel is divided into smaller subsections. The running time, the dwell time and signalized intersection delay time in each of these subsections are estimated. Let $t(k)$, $x(k)$, $d(k)$ and $I(k)$ represent the total travel time, running time, dwell time and signalized intersection delay time of a bus for covering the k^{th} subsection. The running time, the dwell and intersection delay time are estimated using schemes presented below and the sum of the estimated values of these three variables for a particular subsection is taken to be the total travel time for that subsection. The bus traveling track and the relationship of bus arrival Time, link travel time and signalized intersection delay time as shown in Fig. 2.

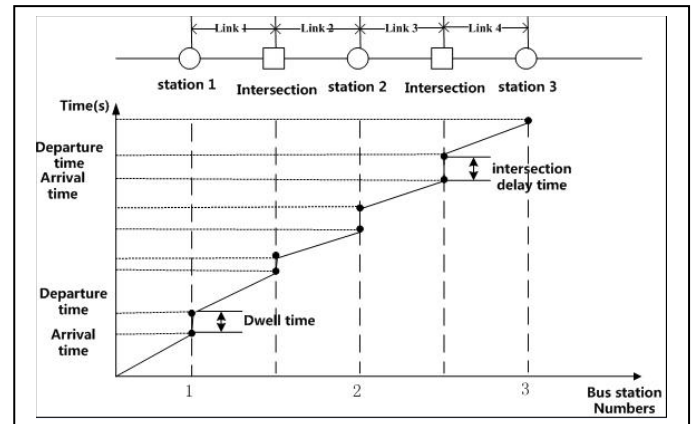


Figure 2. The bus traveling track

- Estimation of running time

The average speed of traffic flow of road network is got from the floating bus in 5 minutes update cycle. So the average travel time of each link of the path can be computed. The bus positions are got through the GPS information. The prediction travel time can be obtained by the summation of the average travel time of each link. The prediction travel time doesn't contain dwell time in the station and the intersection delay time. So this part time is needed to be compensated.

It was assumed that the evolution of running time between the various subsections is governed by

$$T_i = \sum_{i=0}^N \frac{L_i}{V_i} \quad (5)$$

Where T_i is the running time taken for covering the N links, L_i is the length of the i^{th} link. V_i is the average travel time of the i^{th} link.

- Estimation of dwell time

In this article, only the dwell time arising from the stoppage of a bus at a bus stop is considered. It was assumed that the dwell time in the subsection are related through

$$T_h = \frac{\sum_{k=1}^N (T_{dk} - T_{ak})}{N} \quad (6)$$

$$T_i = \frac{\sum_{w=1}^M T_{hw}}{N} \quad (7)$$

Where T_h is the average dwell time of one hour of i^{th} station, T_{dk} is the departure time for the K^{th} bus, T_{ak} is arrival time for the K^{th} bus, N is the number of bus of one line passed this stop for an hour. T_i is the average speed of this period of the day of a week for M weeks, T_{hw} is the W^{th} week. The arrival time, departure time for bus has been predicted.

The dwell time was assumed to be governed by

$$T_w = \sum_{i=j}^k T_i \quad (8)$$

Where T_w is overall dwell time from the current bus location to the destination site.

- Estimation of intersection delay time

The delay time the intersection leads to should be obtained in real time for the Bus Arrival Services System. Historical trend library was applied to supply delay intersection contributes to as a result of the unforeseeability and obvious historical rules for the delay intersection results in.

Intersection delay database is stored in Data cube which the Bus Arrival Services System can acquire delay time in any time interval and any week day.

The prediction intersection delay time can be obtained by the summation of the average delay time of each intersection.

It was assumed that the evolution of running time between the various subsections is governed by

$$T_d = \sum_{j=0}^k T_j \quad (9)$$

Where T_d is intersection delay time from the current bus location to the destination site, T_j is the average delay time of the j^{th} intersection.

B. Long-distance prediction

There are much longer distance and unforeseen conditions in long-distance bus arrival time prediction. For instance, the whole journey will take 3 hours in 944 bus line in which the bus road condition can fall from peak to trough. Therefore, it is unexpected to calculate the bus arrival time simply depending on the road condition.

The model applied for long-distance prediction of bus arrival time is the way based on both real-time road condition and historical rules for bus arrival.

The bus travel time in every stations pair (for instance, 2nd station to 7th station) is modeled. Data cube split by time interval feature and week feature is applied in storing historical data. In research, the bus travel time in adjacent stations doesn't have obviously rule. However, stations apart from multiple stations have stable and obviously rule in bus arrival time. Short forecast will provide Bus Travel time within 3 stations called short-distance bus travel time (STT). The long-distance bus arrival time (LBAT):

$$LBAT = STT + Time(startStation + 2, endstation) \quad (10)$$

C. Virtual prediction

The BAT services system will suffer for its strongly depending on the GPS data. Once GPS services System breaks down and the GPS information is loss, BAT services system won't publish service at all. The GPS information quality is 80% for all the bus lines- that means there's one GPS Data lost every 5 data. In consequence, this paper proposes a way to predict according to virtual GPS point.

For every online bus, a real-time chart is maintained in memory. The chart stores the bus online information (BOI). The structure of BOI can be described as follows:

$$BOI = \left\{ \begin{array}{l} boi | boi = (vehicleId, TimeStamp, routeName, \\ distancepredict, timepredict, stationNum) \end{array} \right\} \quad (11)$$

For every GPS online bus, 5 BOI are maintained in the coach. Supposing that bus GPS information is lost, the model that handle lost GPS can be described as follows:

- GPS information Lost and LBOI is not in station

Actually, the GPS information for one bus is transmit to BAT services system in a transfer period T . Visual GPS information (VGI) is applied to support bus arrival time and distance prediction due to the lost GPS existing. VGI can be described as follows:

$$VGI = \left\{ \begin{array}{l} vgi | vgi = (vehicleId, TimeStamp, \\ routeName, distancepredict, timepredict) \end{array} \right\} \quad (12)$$

Every 20 seconds, a new GPS data coming, the BOI chart can be detected for the bus which lost its GPS data. VGI can be constructed to complement the BOI chart. For one bus,

$$VGI.timepredict = LBOI(last BOI for the bus).timepredict - 20s \quad (13)$$

Based the LBOI's predicted distance LD and current road information. T_k is the road information for the link. L_k is the

length of the k^{th} link. So the distance D the bus cover in the 20 seconds is :

$$\sum_{k=0}^N \frac{L_k}{V_k} = 20s \quad (14)$$

$$D = \sum_{k=0}^N L_k \quad (15)$$

The predicted distance of VGI can be calculated:

$$VGI.distancepredicted = LD - D \quad (16)$$

However, the predicted time and distance for the VGI calculated for the first step may be negative. If the predicted distance is negative, it illustrate that the bus has already passed by *stationNum S*. The station Number for VGI need to be relocated. The predicted distance and predicted time should be recomputed.

First, the predicted distance is recomputed. $dis(stationNum S, stationNum S+1)$ means the distance between the station S and Station $S+1$.

$$VGI.distancepredicted = LD - D + dis(stationNum S, stationNum S+1) \quad (17)$$

Second, the predicted time is recomputed. Link j where the bus VGI represent is on is found by the $VGI.distancepredicted$. Link N is where station $S+1$ is on. The $VGI.timepredicted$ time is:

$$VGI.timepredicted = \sum_{i=j}^N \frac{L_i}{V_i} \quad (18)$$

Then if $VGI.timepredicted$ is negative and $VGI.distancepredicted$ is positive, the link j where bus VGI is on need to be relocated and $VGI.timepredicted$ need to be recomputed according to the road condition.

- GPS is lost. LBOI is in station.

In actual bus arrival services operation, a station delay SD_i is maintained in the coach. There is an delay item for any station. Assuming that the distance to station is less than 100, considering for the station delay, VGI is calculated as follows:

Link ls is link where station S is on. Link lsn is where station $S+1$ is on.

$$VGI.distancepredicted = dis(stationNum S, stationNum S+1) \quad (19)$$

$$VGI.timepredicted = SD_s + \sum_{i=ls}^{lsn} \frac{L_i}{V_i} \quad (20)$$

- GPS lost. LBOI is near the final station.

Assuming that the next station for the LBOI is final station but is not in station, Model 1 is applied to predict the bus arrival time and distance. However, the BOI for the bus should be removed if the bus has already entered the final station.

IV. CORROBORATION OF THE ESTIMATION SCHEME

The Mean Absolute Percentage Error(MAPE), Mean Absolute Percentage Error For A Line (MAPEL), Mean Absolute Error (MAE) and the Percent MAPE less than 80% occupy(PMAPE) are used as a measure of estimation accuracy and is calculated using

$$\begin{aligned} MAPE &= \frac{1}{N} \sum_{k=1}^N \frac{|t(k) - t_m(k)|}{t_m(k)} * 100 \\ MAPEL &= \left(\frac{1}{N} \sum_{k=1}^N \frac{|t(k) - t_m(k)|}{t_m(k)} \right) * 100 \\ MAE &= \frac{1}{N} \sum_{k=1}^N |t(k) - t_m(k)| \end{aligned} \quad (21)$$

where $t_m(k)$ is the arrival time of the test buses measured from the field. N is the test times.

The efficacy of the estimation scheme proposed in the previous section was tested using the measured data from the 944 bus line Beijing and the results are presented below.

When the objective site is within three stations, the time prediction based on the point was carried out using the above model. The estimated arrival time and the measured arrival time of the 20 operating buses are compared and the results are presented in Figs. 3. In this case, it's easy to see that the more close the distance to the objective site, the more accurate the value between the predicting time and the measuring time. The maximum error is 50s.

When the bus is far from the bus stop, The time prediction based on the path was carried out using the above model. The estimated arrival time and the measured arrival time of one bus for the 17 bus stations are compared and the results are presented in Figs. 4. And the estimated arrival time and the measured arrival time for 15 times on one station are compared and the results are presented in Figs. 5. It can be observed that the prediction results agree reasonably well with the measured data.

The values of the MAPE for these three ways were obtained to be 13.3 %, 6.8%, 12.9%.

The PMAPE for lines is displayed as the chart:

TABLE I. PMAPE RESULT FOR LINES

Line name	Direction	PMAP	MAE<30s
977	0	95.3%	86.3%
945	1	94.8%	86.1%
988	1	93.2%	85.3%

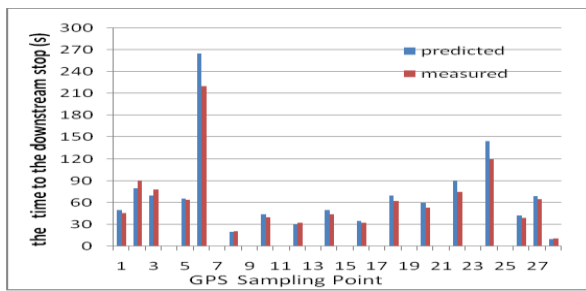


Figure 3. Comparison of the predicted and measured time for the 20 operating buses

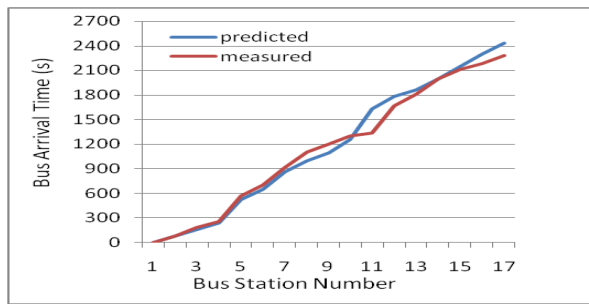


Figure 4. Comparison of the predicted and measured time for the 17 bus station

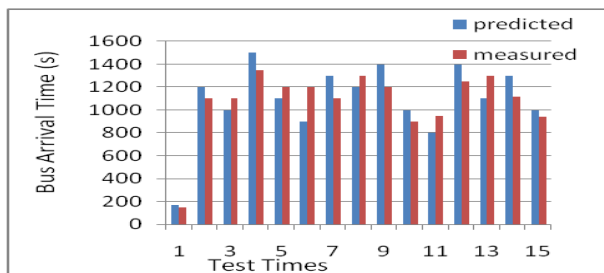


Figure 5. Comparison of the predicted and measured time for 15 times on one station

V. CONCLUDING REMARKS

The comparison of the predicted and measured time reveals that the traffic flow data can better predict link travel time values because it is sensitive to the variations in traffic flow. The historical GPS data and traffic flow data show similar performances in normal traffic conditions, whereas the latter data are superior when there is an abnormal surge in traffic flow. In a comparative sense, the traditional model shows the worst performance since it is not able to capture the variations in travel times.

The traffic flow data based model developed in this study can be regarded as a promising means for the use in advanced traveller information systems where collecting traffic flow is feasible. Among many factors affecting bus travel time, this study uses traffic flow data and schedule adherence to predict bus travel time. The inclusion of other variables such as passenger demand to predict bus travel time might improve the

accuracy of the results. This is a promising direction for future studies.

ACKNOWLEDGMENT

This research was supported by Key transportation energy - saving technology based on vehicle networking and applied research (No.2012AA111903).

REFERENCES

- [1] W. H. Lin and J. Zeng, "An experimental study of real-time bus arrival time prediction with GPS data," *Transp. Res. Rec.*, no. 1666, pp. 101-109, Jan. 1999.
- [2] Z. Wall and D.J. Dailey, "An algorithm for predicting arrival time of mass transit vehicle using automatic vehicle location data," in *Proc. 78th annual meeting, Transportation research Board, Washington, D.C.*, Jan. 1999.
- [3] S. I. J. Chien, Y. Ding, and C. Wei, "Dynamic bus arrival time prediction with artificial neural networks," *J. Transp. ASCE*, Vol. 128, pp. 429-438, Sep/Oct. 2002.
- [4] F. W. Cathey and D. I. Dailey, "A prescription for transit arrival/departure prediction using automatic vehicle location data," *Transportation Research Part C: Emerging Technologies*, Vol. 11, pp.241-264. Jun-Aug. 2003.
- [5] H. Liu, H. J. van Zuylen, H. V. Lint, and M. Salomons, "Urban arterial travel time prediction with state-space neural networks and Kalman filters," *Transportation Research Board, Annual Meeting (CD-ROM)*, Washington, D. C., pp. 99-108, Jan. 2006.
- [6] R. P. S. Padmanaban and L. Vanajakshi, "Estimation of Bus Travel Time Incorporating Dwell Time for APTS Applications," *IEEE Intelligent Vehicles Symposium*. Vol.1, pp.955-959, 2009.
- [7] L. Vanajakshi, S. C. Subramanian, and R. Sivanandan., "Short-term prediction of travel time for indian traffic conditions using buses as probe vehicles," in *Transportation Research Board, the 87th Annual Meeting (CD-ROM)*, Washington, D. C., Jan. 2008.
- [8] Farhan and A. Shalaby and T. Sayed. "Bus Travel Time Prediction Using GPS and APC", *ASCE 7th International Conference on Applications of Advanced Technology in Transportation*, Cambridge, Massachusetts, August 2002.
- [9] J.S. YANG: 'Travel time prediction using the GPS test vehicle and Kalman filtering techniques'. *American Control Conf.*, Portland, OR, USA, 2005, pp. 2128-2133
- [10] R. JEONG, L.R. RILETT: 'Bus arrival time prediction model for real-time applications'. *Transportation Research Record: Journal of the Transportation Research Board*, No.1927, TRB, National Research Council, Washington, D.C., 2005, pp. 195-204
- [11] M. CHOWDHURY, A. SADEK, Y. MA, N. KANHERE, P. BHAVSAR 'Applications of artificial intelligence paradigms to decision support in real-time traffic management'. *Transportation Research Record: Journal of the Transportation Research Board*, No.1968, TRB, National Research Council, Washington, D.C., 2006, pp. 92-98
- [12] M. CHEN, S.I.J. CHIEN: 'Dynamic freeway travel time prediction using probe vehicle data: link based vs. path based'. Presented at the 80th Annual Meeting (CDROM), TRB, National Research Council, Washington, D.C., 2001
- [13] A. SHALABY, A. FARHAN: 'Bus travel time prediction model for dynamic operations control and passenger information systems'. *Transportation Research Board, National Research Council, Washington, D.C.*, 2003, CD-ROM
- [14] S.I.J. CHIEN, C.M. KUCHIPUDI: 'Dynamic travel time prediction with real time and historical data'. Presented at the 81st Annual Meeting (CDROM), TRB, National Research Council, Washington, D.C., 2002

Application Test Process in Product Life Cycle

Oner Tekin
Technology Leader
Netas
Istanbul, Turkey
oner@netas.com.tr

Gulsah Bayram Cetin
R&D ICT Test and Automation
Netas
Istanbul, Turkey
gulsah@netas.com.tr

Abstract—This article explains the phases of application testing process in product life cycle and methodologies of software testing. Software testing is a series of processes begin with requirements step in the early phases of product life cycle and there are many different types of testing methods or techniques used as part of the software testing methodology. The methodology of software deployment chosen depends on the nature of project. Most software development projects involve periodic testing, but some methodologies focus on getting the input from testing earlier in the product life cycle.

Index Terms—Application Test Process, Test Automation

I. INTRODUCTION

Software testing is a series of processes begin with requirements step in the early phases of product life cycle and a part of the Software Quality Assurance (SQA) process. Various stages of testing that occur throughout a project are illustrated in “Fig. 1”. These stages are:

- Requirements
- Test Strategy documentation and reviews
- Test Plan documentation and reviews
- Test Case documentation and reviews
- Test development for Software Test Automation applicable projects
- Test execution
- Result and Defect Reporting
- Final Assessment Testing
- Trial Preparation Testing
- Customer Acceptance Preparation Testing

The software testing techniques could be implemented in two ways - manually or by automation. There are many software testing techniques used as part of the software testing methodology:

- Unit Testing
- Integration Testing
- Feature Verification
- Product Integration Testing
- Product Verification
- Solution Verification
- Regression Testing

- Performance Testing
- Recovery Testing
- Security Testing
- Conformance Testing
- Compatibility Testing

There are numerous methodologies available for developing and testing software. One of the common methodologies is “Waterfall Model”. This model is a sequential development process which allows dividing life cycle into various phases. Iterative models break the cycle into parts and "Waterfall Model" is applied to each part. In "V Model", development and testing takes place at the same time with the same kind of information.

II. APPLICATION TESTING PROCESS

This section explains, stages of testing that occur throughout a project regardless of the testing methodology. Product Life Cycle is a procedural process and software test teams involve at the customer requirements step of software development process.

A. Software Requirements Specification

Software testing is a series of processes begin with requirements step in the early phases of product life cycle. Test architects and/or specialists should take part in requirements stage. Timeline, milestones of sub processes of test cycle, hardware/software requirements for testing, and resources should be identified at this phase. All requirements should be documented properly for further use in “Software Requirements Specification”.

B. Test Strategy Documentation and Reviews

“Test Strategy” describes the high level planning which identifies the resources, equipment, tools, test activities and internal milestones required to meet the program objectives. The strategy defines the holistic product or solution view of the system verification work in a given release. Strategy document is reviewed by Product Line Management, Development, Product Support teams, and Test Managers/Architects.

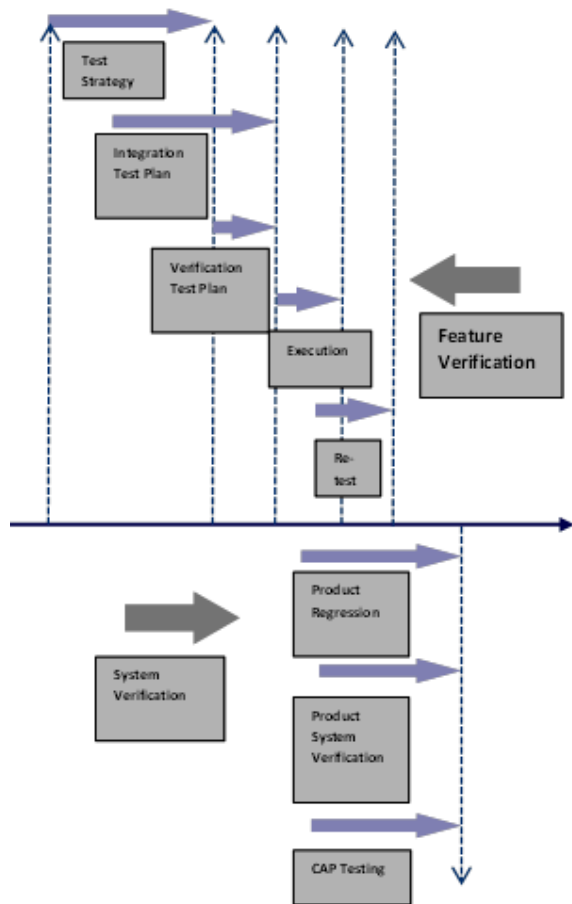


Fig. 1. Test Process

C. Test Plan and Test Case Documentation and Reviews

“Test Plan” describes the low level planning which identifies the detailed resources, equipment, tools, and test activities, etc. Detailed test scenarios are defined in “Test Case” document.

D. Test Development for Software Test Automation

In today’s software projects, test automation becomes an increasingly critical and strategic necessity. Automation Scripts are developed in this phase and updated during execution as necessary.

E. Test Execution, Result and Defect Reporting

Test Cases which have been developed in the Test Activity Planning phase are executed on the target system and/or product. The strategy addresses the mechanics of issue tracking as it interacts with other processes as well as the test execution metrics collection and reporting. Defects are not caused only due to coding errors, but most commonly due to the requirement gaps in the non-functional requirement. A failure is caused due to the deviation between an actual and an expected result.

F. Final Assessment Testing

Defines the work required to perform the final assessment of a product or solution by identifying and executing select test suites upon receipt of the Final Compile loads.

G. Trial Preparation Testing

Focuses on outlining the work requirements associated with product or solution preparation for customer specific trials. It also addresses trial specific planning, testing and tracking.

H. Customer Acceptance Preparation Testing

Defines the workflow associated with the introduction of customer acceptance criteria testing.

III. SOFTWARE TEST MODELS

A. Unit Testing

The primary goal of unit testing is to take, individual component within a subsystem, and isolate it from the remaining components of the code, and determine whether it behaves exactly as you expect. Each component/unit is tested separately before integrating them into modules to test the interfaces between modules. Development teams are responsible for unit testing and it is an example of White-Box testing. White-Box testing assumes that the test specialist could follow the related part of the code, and create the test cases for the potential failures.

The number of defects found in the development is high and the cost to fix those defects is relatively low when comparing to further stage of product life cycle.

B. Integration Testing

Next step is to test the interoperability of the different components of a subsystem which is known as “Integration Testing”. Software Integration Testing covers off network element specific stress testing, regression testing, and search and destroy. Also ensuring that network element interfaces are working correctly, and design teams are responsible for software integration.

Feature Integration Testing verifies the basic feature operation following integration into the release build or integration stream. Tests verify that the feature interactions with other architecture components or activities. Design and/or test teams are responsible for feature integration tests.

There are three common approaches of implementing integration tests:

1) Top-Down Approach

High level units are tested and combined first. This enables early testing of high-level logic.

2) Bottom-Up Approach

In this approach, lower level units are tested and combined first.

3) Umbrella Approach

This involves testing the control flow paths and the functional data. Initially, the inputs for the different functions are combined using the bottom-up approach. The outputs are then combined utilizing the top-down approach.

C. Feature Verification

Feature Verification provides a complete verification of feature operation in a working system for both success path and failure path against the requirements in the Feature Requirements Spec and Feature Technical Spec. Design and/or test teams are responsible for feature verification testing. This type of testing is an example of Black-Box testing. Black-Box testing assumes that the test specialist creates and implements the test cases without any internal knowledge of the unit under test. This approach tests all possible combinations of end-user.

D. Product Integration Testing

Product Integration is a regression testing of earlier release features to ensure components and interfaces are still intact amongst network elements. Design and/or test teams are responsible for product integration testing.

E. Product Verification

At the phase of product verification, the main goal is to ensure the system as a whole including all new features can respond favorably to robustness scenarios, capacity stress tests, reliability testing, extended soaking, end to end upgrade and rollbacks and deployment scenarios. Test teams are responsible for the product verification.

F. Solution Verification

Solution verification verifies specific end-to-end functionality and performance of the solution in a customer simulated environment, where realistic conditions are critical to the tests, and test teams are responsible for the solution verification.

G. Regression Testing

Regression testing verifies the existing behavior of the software is unbroken after new implementations. Regression testing is a good application area for software test automation.

H. Performance Testing

Performance testing verifies that a system performs in terms of responsiveness and stability under a particular workload and meets the specifications claimed by its manufacturer or vendor. Software test automation is applicable to performance testing.

I. Recovery Testing

Recovery testing verifies that how a system or application is able to recover from hardware and software crashes and failures. This type of testing determines the ability to restart the applications after integrity lost. The main focus is to ensure that the functionality continuous after disaster cases.

J. Security Testing

Security testing determines whether a product protects data and maintains functionality of the system properly. Main focus is to protect confidentiality. Vulnerability scanning finds weakness of the system. Penetration testing simulates the malicious attacks to the system.

K. Conformance Testing

Conformance determines whether a product meets the international and industrial standards and interoperates properly.

L. Compatibility Testing

Compatibility testing determines whether a product compatible with other elements of the system or environment. Software test automation is applicable to compatibility testing.

IV. SOFTWARE TESTING METHODOLOGIES

The methodology of software deployment chosen depends on the nature of project. Most software development projects involve periodic testing, but some methodologies focus on getting the input from testing earlier in the product life cycle.

A. Waterfall Model

One of the common methodologies is “Waterfall Model” which is a non-iterative approach. This model is a sequential development process which allows dividing life cycle into various phases. “Waterfall Model” is shown in “Fig. 2”.

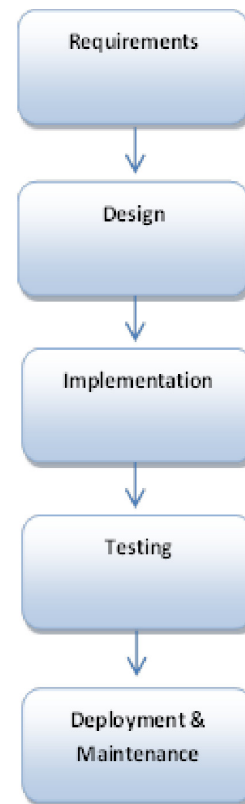


Fig. 2. Waterfall Model

B. V Model

In this approach, both development and test teams drive the process in parallel (It could be visualized forming the letter 'V'). Development and testing takes place at the same time with the same kind of information. "Fig. 3", represents the "V model".

C. Spiral Model

In the spiral development environment, software testing is again described as a continuous improvement process that must be integrated into a rapid application development methodology. Testing as an integrated function prevents development from proceeding without testing. There are a number of cycles of all the sequential steps of the waterfall model. Once the initial cycle gets completed, a thorough analysis and review of the achieved product or output is performed. "Fig. 4", represents the Spiral Model.

D. Agile Model

Agile is a hybrid method which mixes sequential and iterative approaches. The main focus is to create quick, practical and visible outputs. This is a practice-based methodology. Agile development is successful when software test automation is implemented properly. This model relies on the repetition of a very short development cycle.

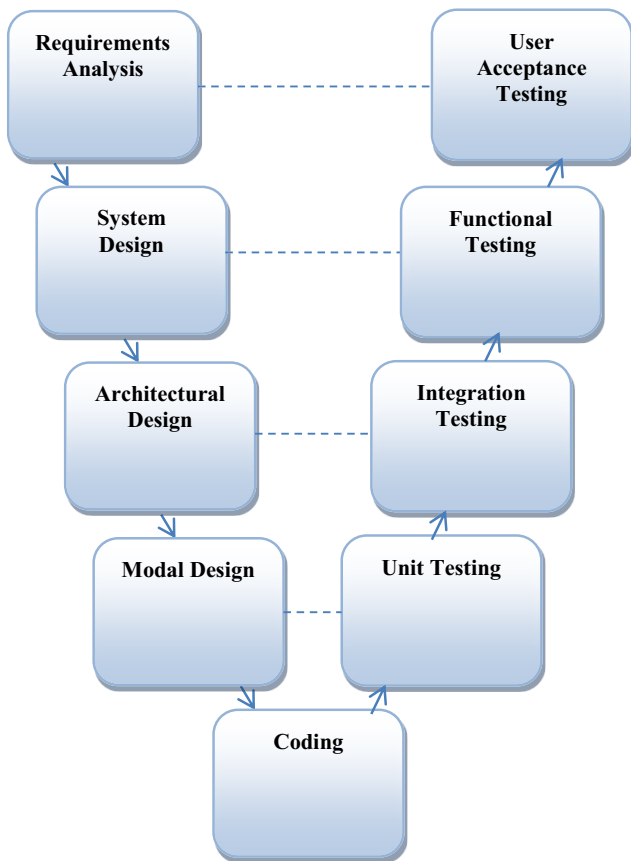


Fig. 3. V Model

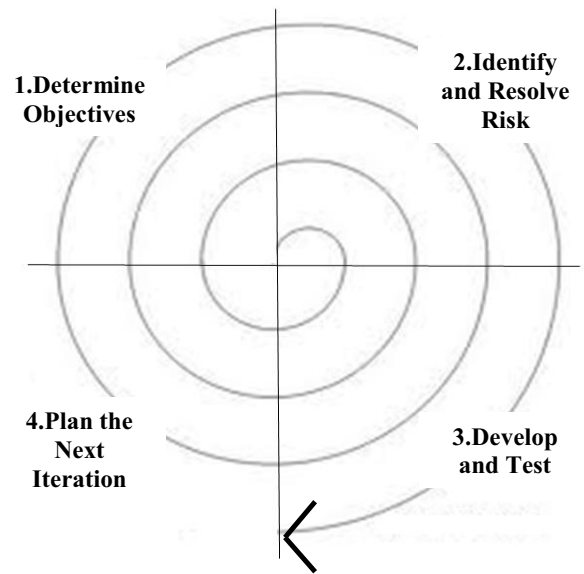


Fig. 4. Spiral Model

"Figure 5" determines Agile methodology. Agile principles are:

- Continuous Delivery
- Quick handling of changing requests
- Quick software delivery
- Involvement of business units and developers throughout the project
- Self-organized teams
- Good communication between the business units
- Technical excellence and good design

There are different methods of Agile software development methodology.

1) Scrum

The basic idea of scrum is, there are several technical and environmental variables changing during the process. These variables are requirements, resource, time, and technology. Scrum can help organizations to achieve better engineering activities, as it has some frequent activities for management to check the system regularly for any kind of deficiencies.

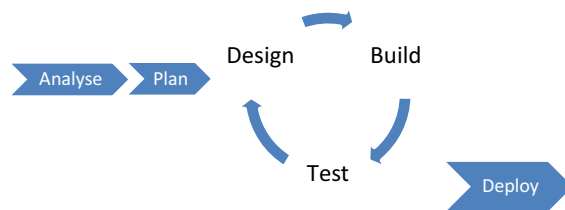


Fig. 5. Agile Model

2) Extreme Programming

Extreme Programming is a collection of different models. The main characteristics are, short iterations with rapid feedback, and small releases, regular participation of the customer, continuous testing and integration, collective code ownership, not a detailed documentation.

3) Feature Driven Development

Feature Driven Development (FDD) is a client-centric, architecture-centric software process. In FDD, testing is done through inspection. In FDD process, unit tests are taken and implemented according to the project and situation.

4) Crystal Clear Method

Crystal Clear method is used for small teams. It attempts to describe a full methodology of the lightest, most habitable kind that will produce good results.

V. SOFTWARE TEST AUTOMATION

Software test automation is a process which uses software for testing a software product. Automation is a kind of development process in this manner. If there are repeatable actions in test scenarios, software test automation is an effective way to implement these test cases.

It is critical to describe the automation test objectives for the first step. Test architects and/or test automation experts are responsible for designing the test framework and architecture. Test specialists must then comply with and contribute to the architecture and design automation scripts. Parameterization of scripts is one of the important steps for software test automation.

While the development is in progress, as soon as the functional description is completed, test plan and detailed test cases are created. Test libraries for common functions and test scripts for base-unrelieved applications could be created, at this point but this is an early phase to create all test scripts. Test architects should start to build test framework. It is an important phase to identify automated, semi-automated, and manual test cases. When the development completed, automation team could start to create test scripts for new activities. These scripts will be updated when the functionality change. Test automation maintenance is an important issue, and automation teams should handle any changes in related product. Automation provides many advantages in software testing:

A. Speed

Testing process is expedited, since a program naturally works quicker than the pace of a human tester. Many automated testing tools can replicate the activity of a large number of users using a single computer. Therefore, load/stress testing using automated methods require only a fraction of the computer hardware that would be necessary to complete a manual test. Parallel execution reduces the execution effort and time.

B. Cost Effectiveness

The cost of performing manual testing is prohibitive when compared to automated methods. Automation is not time

dependent, so the scheduling methods allow us to execute the automated test suites any time.

C. Reliability

In automation, human factor is virtually eliminated, giving less consideration for subjectivity and possible margins of error. Full-featured automated testing systems also produce convenient test reporting and analysis. These reports provide a standardized measure of test status and results, thus allowing more accurate interpretation of testing outcomes. Manual methods require the user to self-document test procedures and test results.

D. Reusability

Automated tests are reusable on different versions of an application.

E. Comprehensive

A suite of tests that covers every feature in the application could be build. The productivity gains delivered by automated testing allow and encourage organization to test more often and more completely. Greater application test coverage also reduces the risk if exposing users to malfunctioning or non-compliant software.

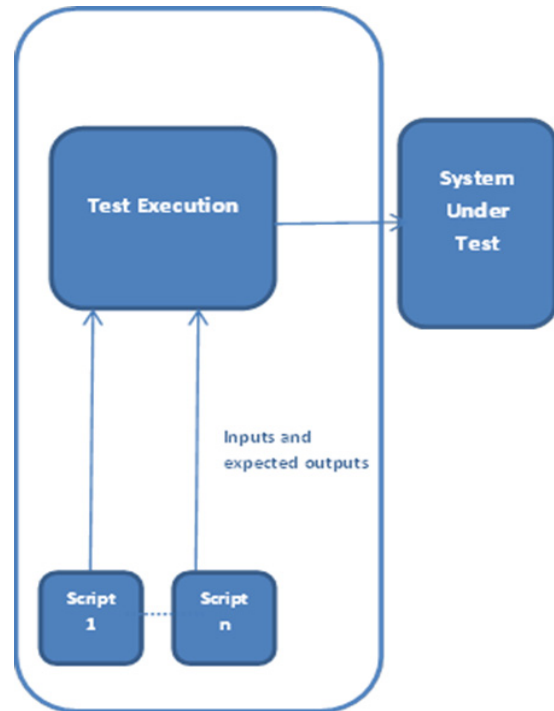


Fig. 6. Test Automation Framework

F. Repeatable Testing

Automation allows the testing organization to perform consistent and repeatable test. When applications need to be deployed across different hardware or software platforms, tests

can be created and repeated on target platforms to ensure that new platforms operate consistently.

The methods that are employed to carry out testing remain repetitious throughout the development life cycle.

G. Programmable Testing

In automation projects, more sophisticated test suites that bring out hidden information from the application could be created.

H. Good Result Analysis and Resulting

Automated test results could be tracked easily, and all iterations of execution are reported.

VI. CONCLUSION

Successful software testing projects depends on a standard testing process, good testing objectives and the right team roles, skills and tools. The automation test team needs to have knowledge of testing, programming, and automation tool as well. In today's system, rapid deployments and quick responses to the customers are essential. Software development and test processes during a product life cycle are adapted to these requirements.

REFERENCES

- [1] Koomen, T. and M. Pol, "Test Process Improvement: A practical step-by-step guide to structured testing," Addison-Wesley, 1999
- [2] Copeland, L., "A Practitioner's Guide to Software Test Design," STQE Publishing, 2004
- [3] Graham, D. and M. Fewster, "Experiences of Test Automation: Case Studies of Software Test Automation," Pearson Education, 2012
- [4] Hetzel, B., "The Complete Guide to Software Testing," Wiley, 1993
- [5] Galin, D., "Software Quality Assurance: From Theory to Implementation," Addison-Wesley, 2003
- [6] Dustin, E., Rashkal J. and J. Paul, "Automated Software Testing: Introduction, Management, and Performance: Introduction, Management, and Performance," Addison-Wesley Professional, 1999
- [7] Venkatasubramanian, A. and V. Vinoline, "Software Test Factory (A proposal of a process model to create a Test Factory)," International Journal of Computational Intelligence Techniques, vol. 1, no.1 2010, pp.14-19.
- [8] Boehm, B., "A Spiral Model of Software Development and Enhancement," Computer, vol.21, no.5 1988, pp.61-72.
- [9] Nawaz, A and M. Masood, "Software Testing Process In Agile Development," Master Thesis, Dept. of Computer Science, Blekinge Tekniska Hogskola, 2008.

Distributed File System as a basis of Data-Intensive Computing

Assoc. Prof. Abzettin Adamov
Chair, Computer Engineering Department,
Qafqaz University, Baku, Azerbaijan
aadamov@qu.edu.az

Abstract - The extremely fast grow of Internet Services, Web and Mobile Applications and advance of the related Pervasive, Ubiquity and Cloud Computing concepts have stimulated production of tremendous amounts of data available online. Event with the power of today's modern computers it still big challenge for business and government organizations to manage, search, analyze, and visualize this vast amount of data as information. Data-Intensive computing which is intended to address this problems become quite intense during the last few years yielding strong results.

Data intensive computing framework is a complex system which includes hardware, software, communications, and Distributed File System (DFS) architecture. This paper is giving comprehensive information on how distributed file system supports this approach of processing extra-large volumes of data. It is definitely expected that this work will contribute to future research on similar and related topics as spin off from this study.

Keywords: Data-Intensive Computing, Distributed File System, Fault-Tolerant System, GFS, HDFS

I. WHY BIG DATA BECAME SO BIG?

According to a recent IDC's Digital Universe Study the global volume of digital data increased by 62% between 2008 and 2009 to approximately 800,000 petabytes (PB). The report also noted that the rapidly expanding "Digital Universe" is expected to grow to 1.2 million PB, or 1.2 zettabytes (ZB) in 2010, 1.8 zettabyte in 2011, this year it is expected to grow up to 2.7 zettabytes (ZB) and reach 35 ZB by 2020. [1]

Data sets also grow in size because they are increasingly being gathered by ubiquitous information-sensing mobile devices, aerial sensory technologies (remote sensing), software logs, cameras, microphones, Radio-frequency identification readers, and wireless sensor networks.

Over 90% of this information will be unstructured, that is in the form which does not have predefined data model or information which is not within RDBMS. Examples of Unstructured Data may include books, journals, documents, audio, video, textual and binary files, the body of an e-mail message (email headers are structured information, since most of them have strong format and meaning), Web-page, etc. Generally, unstructured data is useless unless applying data mining or data extraction techniques in order to extract structured data. Two questions may arise naturally in connection with these forecasts: Why we produce this data if

never consume it? What do we need in order to use this extra-large amount of data and benefit from it [2]?

II. DISTRIBUTED FILE SYSTEMS (DFS)

Because shared files are widely distributed across networks, there are growing problems of keeping users connected to the data they need. The distributed file system provides a mechanism of providing logical views of directories and files, regardless of where those files physically reside in the network. Fault tolerance of network storage resources is one of the most important characteristic of DFS [3].

A. Why is Distributed File System Useful

- Provide common view of centralized file system, but distributed over network
- Ability to open & update any file on any machine on network
- All of synchronization issues and capabilities of shared local files
- Data sharing of multiple users
- User mobility
- Location transparency
- Location independence
- Backups and centralized management

B. Motivation for DFS Implementation

- Need for a scalable DFS
- Large distributed data-intensive applications
- High data processing needs
- Performance, Reliability, Scalability and Availability
- Acute need for the systems with fault-tolerant capabilities

III. GOOGLE FILE SYSTEM

Google as one of the biggest Internet companies invented Google invented its own distributed file system named Google File System (GFS). Since the company is using this platform for all cloud services. According to GSF platform all application data is stored in huge files, so that data collecting

by hundreds distributed machines may be stored in the same file.

Google's engineers designed GFS with clear intention to turn mass of cheap servers into distributed, reliable, fault-tolerant and scalable system which enables to store hundreds of terabytes of data and provide remote access to this data by thousands of users and applications simultaneously at high speed. [4]

Like RAID 5 spreads data across multiple discs in order to avoid data losses, similarly GFS distributes same-size files (chunks) across network. Chunks which are parts of the same file, theoretically may be replicated in cluster servers or even in geographically dispersed server nodes (see Fig. 1).

GFS is designed especially for enabling large data reads and transfers at extremely high speed [5].

A GFS cluster typically consists of a single master, multiple chunkservers and multiple clients as it is shown in figure 1. A file is divided into fixed-size chunks (64MB) and these chunks are distributed over the chunkservers. Each chunk is identified by a unique 64-bit chunk handle assigned by Master while creating the chunk. Each chunk is replicate on at least 3 (can be changed) other chunkservers to increase the reliability of the system. [6]

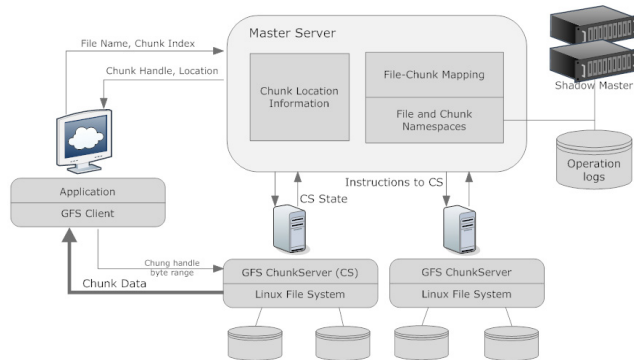


Fig. 1. Google File System Architecture.

IV. HADOOP DISTRIBUTED FILE SYSTEM (HDFS)

HDFS is distributed file system inspired by and similar to GFS. It is scalable to thousands of nodes which run on commodity hardware, so that HDFS assumes failures (both hardware and software) are common, targeted towards small number of very large files, write once and read multiple times. It has many similarities with existing distributed file systems. However, the differences from other distributed file systems are significant. HDFS is highly fault-tolerant and is designed to be deployed on low-cost hardware. HDFS provides high throughput access to application data and is suitable for applications that have large data sets.

Hadoop files are composed of set of fixed-size blocks (typically 64MB). Each block is stored as a separate file on the file system (see Fig. 2).

HDFS's placement policy meets the objectives of load balancing, fast access, fault tolerance. Each block is copying

three times into different datanodes, and this process is also called "Block Level Replication"

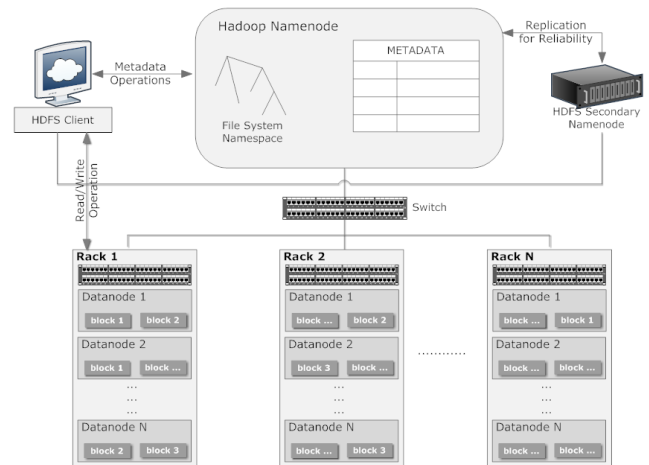


Fig. 2. Hadoop Distributed File System Architecture.

HDFS has two main types of Nodes: NameNode, and DataNodes. HDFS stores all filesystem metadata information on NameNode and actual data in DataNodes. NameNode is constantly checking the state of DataNodes. Everytime a client application wants to read or write data, it sends a message to the NameNode and the NameNode checks where the data should be read from or written to. After receiving appropriate location the client application reads/writes data directly to the DataNode.

A. Cluster Rebalancing

Another essential feature of the HDFS architecture is compatibility with data rebalancing schemes. A scheme might automatically move data from one DataNode to another if the free space on a DataNode falls below a certain threshold. In the event of a sudden high demand for a particular file, a scheme might dynamically create additional replicas and rebalance other data in the cluster. [7]

B. What is not typical for HDFS?

- HDFS is not a POSIX file system
- HDFS is not designed for low latency access to a huge number of small files
- HDFS is not a platform for relational database and does not support transactions
- HDFS is not focused on security, encryption or multitenancy

CONCLUSION

Distributed File System is essential component of any Large-scale Data Processing or Data-Intensive

Computing System. Although, there are many implementations of DFS, but GFS and HDFS are the most successful platforms that became significant and vital component of digital infrastructure of biggest Internet players like Google, Yahoo, IBM, Amazon, etc. Even these DFSs provide relatively reliable, scalable and fault-tolerant platforms, there is still a need for challenge of new approaches and new ways of solving Data-Intensive Computing problems.

REFERENCES

- [1] Digital Universe Study, IDC analysts, June 2011, URL: <http://www.emc.com/leadership/programs/digital-universe.htm>
- [2] Gokhale, M.; Cohen, J.; Yoo, A.; Miller, W.M.; Jacob, A.; Ulmer, C.; Pearce, R.; , "Hardware Technologies for High-Performance Data-Intensive Computing," *Computer* , vol.41, no.4, pp.60-68, April 2008, doi: 10.1109/MC.2008.125, URL: <http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=4488252&isnumber=4488232>
- [3] Step-by-Step Guide to Distributed File System, Microsoft, URL: <http://technet.microsoft.com/en-us/library/bb727150.aspx>
- [4] Sakr, S.; Liu, A.; Batista, D.M.; Alomari, M.; , "A Survey of Large Scale Data Management Approaches in Cloud Environments," *Communications Surveys & Tutorials, IEEE* , vol.13, no.3, pp.311-336, Third Quarter 2011, doi: 10.1109/SURV.2011.032211.00087, URL: <http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=5742778&isnumber=6026692>
- [5] Sean Gallagher, The Great Disk Drive in the Sky: How Web giants store big—and we mean big—data, Jan 27 2012, URL: <http://arstechnica.com/business/2012/01/the-big-disk-drive-in-the-sky-how-the-giants-of-the-web-store-big-data/>
- [6] Sanjay Ghemawat, Howard Gobioff, The Google File System, URL: <http://research.google.com/archive/gfs.html>, October, 2003
- [7] Robert Chansler, Hairong Kuang, Sanjay Radia, The Hadoop Distributed File System, URL: <http://www.aosabook.org/en/hdfs.html>

DIRECT ROBUST NON-NEGATIVE MATRIX FACTORIZATION AND ITS APPLICATION ON IMAGE PROCESSING

Bin Shen, Zhanibek Datbayev
 Department of Computer Science
 Purdue University
 West Lafayette, IN, 47907, USA

Olzhas Makhambetov
 Department of Computer Science
 Center for Energy Research
 Astana, Kazakhstan

Abstract— In real applications of image processing, we frequently face outliers, which cannot be simply treated as Gaussian noise. Nonnegative Matrix Factorization (NMF) is a popular method in image processing for its good performance and elegant theoretical interpretation, however, traditional NMF is not robust enough to outliers. To robustify NMF algorithm, here we present Direct Robust Nonnegative Matrix Factorization (DRNMF) for image denoising based on the assumptions that the ground truth data is of low rank and the outliers are sparse. This method explicitly models the outliers in the data, and the sparsity of the outliers is controlled by L_0 norm. The experiments show that DRNMF can accurately localize the outliers, and outperforms traditional NMF in image denoising.

Keywords—Outlier removal, sparse error, NMF, image denoising, robust NMF, nonnegative representation

I. INTRODUCTION

Among lots of data processing algorithms, Nonnegative Matrix Factorization (NMF) [1, 2] is famous for its good performance in real applications and good theoretical interpretation. NMF is able to decompose a nonnegative matrix into a product form of two nonnegative ones. An object is represented as a nonnegative combination of nonnegative components. Nonnegative matrix factorization is distinguished from the other methods such as PCA, because of the nonnegativity constraint which allows only non-subtractive combinations. Due to the same constraint, the learned basis and coefficients will naturally have sparsity.

The mathematical definition of NMF is as follows. Given a matrix $X \in R^{m \times n}$. It aims to factorize this matrix in the form of the product of two nonnegative matrices $U \in R^{m \times p}$ and $V \in R^{p \times n}$ by minimizing the objective function:

$$O = \|X - UV\|_F^2 \quad (1)$$

Because this function is not joint convex with respect to both U and V , a local optimum is searched instead of the global optimum. In order to attain the goal of minimizing the objective function, two iterative multiplicative updating rules are proposed in [2] as follows.

$$U_{ij} = U_{ij} \frac{(XV^T)_{ij}}{(UVV^T)_{ij}} \quad (2)$$

$$V_{ij} = V_{ij} \frac{(U^T X)_{ij}}{(U^T UV)_{ij}} \quad (3)$$

NMF is a powerful method to analyze real data according to many researchers' works. Because of its good experimental performance and easy interpretation, developing variants of NMF to adapt it to new settings has attracted a lot of researchers' attention. [3] and [4] propose sparse NMF, which controls the sparsity of the two resulting matrices. [5] extends the NMF to incomplete data, and uses it to learn from incomplete rating matrix. [6, 7] and [8] consider the underlying manifold structure of data samples, and propose NMF on a single manifold. Later, [9] considers the multiple manifold case. More variants of NMF is still being developed, and it is one of the most popular techniques these days.

II. DIRECT ROBUST NONNEGATIVE MATRIX FACTORIZATION FOR LEARNING FROM NOISY DATA

Nonnegative matrix factorization is a powerful technique to describe real data, since a lot of real data are nonnegative. For example, text documents are usually described as a nonnegative matrix; images are also represented as a nonnegative matrix. Nonnegative matrix factorization in formulation (1) assumes the noise of the observed data is drawn from Gaussian distribution. However, in some real situation, we may have some outliers, which are of large value and cannot be modeled as noise drawn from Gaussian distribution. For example, the occlusion on human faces or the pepper salt noise in images. In the framework of PCA, [10] proposed a robust version to address this kind of noise. To handle this kind of noise in the framework of NMF, we propose the Direct Robust Nonnegative Matrix Factorization (DRNMF) algorithm.

A. Direct Robust Nonnegative Matrix Factorization

Traditional NMF tries to minimize the least square error, and the underlying assumption is that the noise of the data is drawn from Gaussian distribution. So the traditional NMF is not robust with respect to outliers, which are usually large value noises. Some recent work tries to deal with outliers [5]. They usually assume positions of the corruption are given ahead, and then ignore the corresponding data entries. However, it is unrealistic to assume that the positions of

corruption are known ahead in many real world applications. Here we propose a Direct Robust Nonnegative Matrix Factorization (DRNMF), which aims to deal with the partial corruption without outlier positions given in the framework of NMF. DRNMF directly optimizes the objective function regularized by L_0 norm of the outliers rather than an approximation of L_0 norm, which is adopted by other approaches [11, 12, 13]. We follow the same notation as Nonnegative Matrix Factorization. Given a matrix $X \in R^{m \times n}$. It aims to factorize this matrix in the form of the summation of an error matrix $E \in R^{m \times n}$, which denotes the outliers, and the product of two nonnegative matrices $U \in R^{m \times p}$ and $V \in R^{p \times n}$ by minimizing the objective function.

B. Formulation

In this formulation, a parameter is employed to make a tradeoff between the reconstruction error and the sparsity of E .

$$O_{DRNMF} = \|X - UV - E\|_F^2 + \lambda \|E\|_0 \quad (4)$$

where we constraint that $X - E \geq 0$. This means the original data should be nonnegative. Note here, the components of E can be either positive or negative. The λ controls how heavy the noise is.

C. Optimization

Now we want to minimize O_{DRNMF} with respect to U, V, E . The objective function O_{DRNMF} is not convex with respect to U, V and E jointly. It is difficult to find a global optimum, similar with traditional NMF, we aim to find a local optimum for this objective function by iteratively updating U, V and E . When E and U are given, V is updated to decrease the objective function.

$$O_{DRNMF} = \|X - UV - E\|_F^2 + \lambda \|E\|_0 \quad (5)$$

$$= \|(X - E) - UV\|_F^2 + \text{constant}$$

Take $X - E$ as the input data, and the standard NMF updating rule of V in equation (2) can be applied here with X replaced by $X - E$. Similarly, when optimizing with respect to U with E and V fixed, the updating rule in equation (3) can be applied.

When U and V are fixed, E is updated to decrease the objective function. The updating rule should be

$$E_{ij} = \begin{cases} (X - UV)_{ij}, & |(X - UV)_{ij}| > \sqrt{\lambda} \\ 0, & \text{otherwise} \end{cases}$$

After the updating above, the constraint $X - E \geq 0$ always holds because both U and V are nonnegative.

The optimization of DRNMF is summarized below:

Algorithm 1 Optimization for DRNMF

Step 1 - Initialization:

Set all the elements of E to 0.

Initialize U, V to random nonnegative matrices.

Step 2 - Fix E , optimize with respect to U, V

Iteratively do the following:

$$U_{ij} = U_{ij} \frac{((X - E)V^T)_{ij}}{(UVV^T)_{ij}} \quad (6)$$

$$V_{ij} = V_{ij} \frac{(U^T(X - E))_{ij}}{(U^TUV)_{ij}} \quad (7)$$

Step 3 - Fix U, V , optimize with respect to E

$$E_{ij} = \begin{cases} (X - UV)_{ij}, & |(X - UV)_{ij}| > \sqrt{\lambda} \\ 0, & \text{otherwise} \end{cases}$$

Step 4 - Check convergence. If it converges, return U, V , and E ; otherwise, go back to step 2.

III. EXPERIMENTAL RESULTS

First we test DRNMF algorithm on numerical synthetic data to show the basic property. Next, we test it on a real application: image denoising, which proves that DRNMF can accurately model the outliers and outperforms traditional NMF.

A. Numerical Results on Synthetic Data

Let us consider the matrix X

$$\begin{pmatrix} 0.0333 & 0.0667 & 0.1000 \\ 0.1333 & 0.1667 & 0.3333 \\ 0.2333 & 1.0000 & 0.3000 \end{pmatrix}$$

Note that an approximation matrix \hat{X}

$$\begin{pmatrix} 0.0333 & 0.0667 & 0.1000 \\ 0.1333 & 0.1667 & 0.2000 \\ 0.2333 & 0.2667 & 0.3000 \end{pmatrix}$$

is of rank 1. And the \hat{X} and X is only different in position (2, 3) and (3, 2). So, if we want to decompose X into the product of two low rank (rank 1) matrices to get the approximation \hat{X} and a sparse error matrix E , which have nonzero entries only at positions (2, 3) and (3, 2).

NMF and DRNMF are applied on this X . The parameter p is set to 1, since we want to find a rank 1 approximation matrix. The parameter λ in DRNMF is set to 0.005.

For NMF, we get the approximation $UV =$

$$\begin{pmatrix} 0.0243 & 0.0936 & 0.0358 \\ 0.0704 & 0.2709 & 0.1037 \\ 0.2518 & 0.9682 & 0.3705 \end{pmatrix}$$

The reconstruction error is: $\|X - UV\|_F = 0.2807$.

For DRNMF, we get the approximation $UV =$

$$\begin{pmatrix} 0.0559 & 0.0757 & 0.0736 \\ 0.1259 & 0.1750 & 0.1658 \\ 0.2319 & 0.3141 & 0.3054 \end{pmatrix}$$

DRNMF also computes E , which is equal to

$$\begin{pmatrix} 0.0000 & 0.0000 & 0.0000 \\ 0.0000 & 0.0000 & 0.1675 \\ 0.0000 & 0.6859 & 0.0000 \end{pmatrix}$$

The nonzero entries of the estimated E indicate there are outliers in these positions. This E indicates that outliers are in position (2, 3) and (3, 2), which is consistent with our observation above.

Also, the reconstruction error is: $\|X - UV - E\|_F = 0.0373$, which means that this approximation is much better than the NMF approximation.

Comparing NMF and DRNMF, NMF models noise as Gaussian distribution, which results in poor approximation when outliers present; while DRNMF can model both Gaussian noise and sparse large errors, specifically, DRNMF is able to locate the outliers automatically.

B. Image Denoising Using Direct Robust Nonnegative Matrix Factorization

Here image denoising experiments are conducted to evaluate the proposed DRNMF algorithm. The denoising experiments are conducted on the Berkeley Segmentation Dataset, which was downloaded from (<http://www.eecs.berkeley.edu/Research/Projects/CS/vision/bsds/>).

Pepper salt noise are put on clean image to generate a noisy (polluted) image. Then we want to use the DRNMF/NMF algorithms to recover the original clean image. The image denoising algorithm using DRNMF/NMF is presented below:

Algorithm 2 Image Denoising Algorithm Using DRNMF/NMF

Input: Noisy image

Prepare noisy matrix X from the noisy image.

-Decompose the image into a large set of patches.

-Use a vector to represent each patch.

-Put all these vectors together, we have a matrix X .

Use DRNMF/NMF, calculate U and V .

Get the denoised matrix $\hat{X} = UV$

Construct the denoised image from \hat{X}

In these experiments, the parameters are fixed. $p = 10$ for both NMF and DRNMF, and λ for DRNMF is set to 0.1. Also, the image patches are of size 7×7 , so each vector represents a patch is of dimension 147, which is equal to $7 \times 7 \times 3$. Note there are three values (R , G , B) for each pixels. We sample a patch in the noisy image at the step of 3.

In each experiment, there are four images shown: the original image, the polluted image, the denoised image by NMF, and the denoised image by the proposed DRNMF.

Take the wolf image for example. The original image, noisy image, denoised image by NMF and denoised image by DRNMF are all shown in Fig. 1. From the result we can easily find that DRNMF is able to do the denoising task much better, while NMF blurs the image when pepper salt noise is present.

Fig. 2 and Fig. 3 show more denoising experimental results.

IV. CONCLUSION AND DISCUSSION

In this paper, we propose DRNMF, which is able to address outliers in the framework of NMF. It employs L_0 norm as the measure of sparsity of outliers, and directly optimizes the regularized objective function without approximation by L_1 norm. Experimental results on both synthetic data and real image data show that DRNMF is able to find the exact positions of large sparse errors and it outperforms traditional NMF in the task of image denoising when pepper salt noise, which is pixel outlier, is present. For future research, automatic adaption to the level of noise, i.e. determining the parameter λ , seems to be interesting and useful in real world applications.

REFERENCES

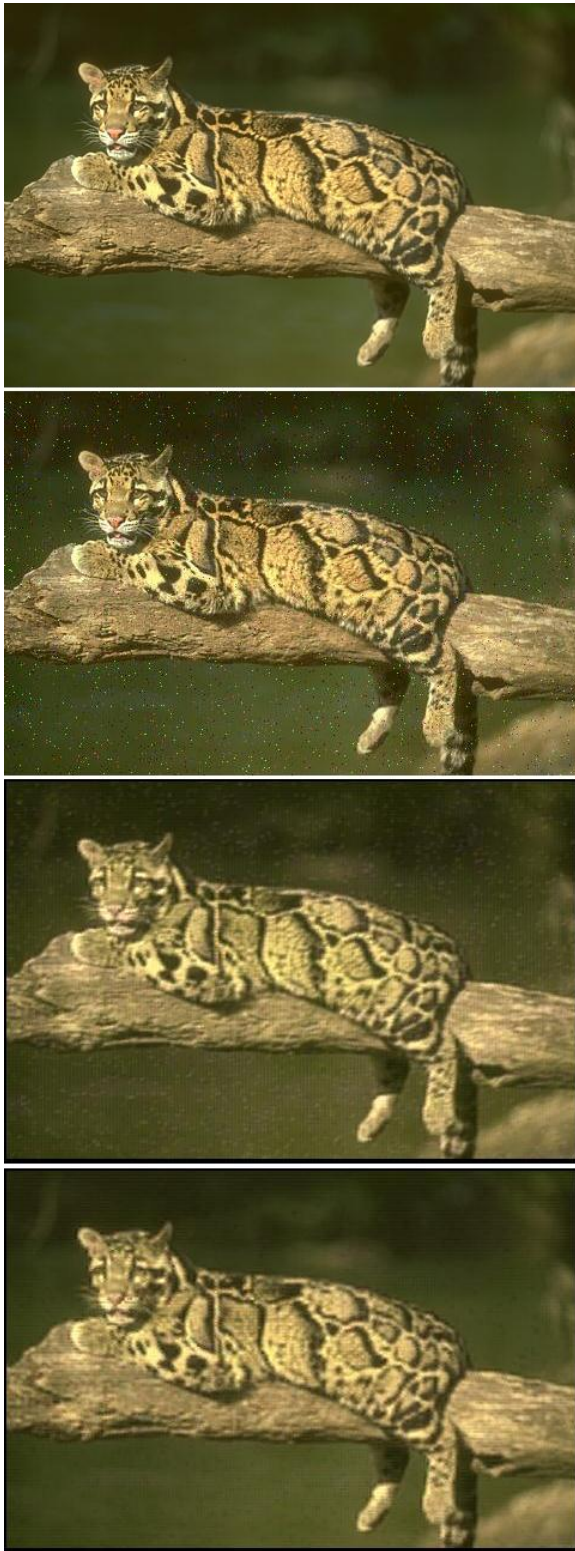
- [1] Daniel D. Lee and H. Sebastian Seung, "Learning the parts of objects by non-negative matrix factorization," *Nature*, vol. 401, no. 6755, pp. 788–791, 1999.
- [2] Daniel D. Lee and H. Sebastian Seung, "Algorithms for non-negative matrix factorization," in *NIPS*. 2001, pp. 556–562, MIT Press.
- [3] P.O. Hoyer, "Non-negative sparse coding," in *Neural Networks for Signal Processing, 2002. Proceedings of the 2002 12th IEEE Workshop on*, 2002, pp. 557–565.
- [4] Jingu Kim and Haesun Park, "Sparse nonnegative matrix factorization for clustering," in *CSE Technical Reports*. Georgia Institute of Technology, 2008.
- [5] Sheng Zhang, Weihong Wang, James Ford, and Fillia Makedon, "Learning from incomplete ratings using non-negative matrix factorization," in *SDM*, 2006.
- [6] Deng Cai, Xiaofei He, Xuanhui Wang, Hujun Bao, and Jiawei Han, "Locality preserving nonnegative matrix factorization," in *IJCAI'09: Proceedings of the 21st international joint conference on Artificial intelligence*, San Francisco, CA, USA, 2009, pp. 1010–1015, Morgan Kaufmann Publishers Inc.
- [7] Deng Cai, Xiaofei He, Jiawei Han, and T.S. Huang, "Graph regularized nonnegative matrix factorization for data representation," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 33, no 8, pp. 1548–1560, Aug.
- [8] Quanquan Gu and Jie Zhou, "Neighborhood preserving nonnegative matrix factorization," in *The 20th British Machine Vision Conference, 2009*.
- [9] Bin Shen and Luo Si, "Nonnegative matrix factorization clustering on multiple manifolds," in *Proceedings of the Twenty-Fourth AAAI Conference on Artificial Intelligence*. 2010, pp. 575–580, AAAI Press.
- [10] Yigang Peng John Wright and Yi Ma, "Robust principal component analysis: Exact recovery of corrupted lowrank matrices by convex optimization," in *NIPS*. 2009, MIT Press.



Figure. 1. Denoise Result of Wolf Image-From top to bottom: Original image; noisy image; denoised by NMF; denoised by DRNMF



Figure. 2. Denoise Result 2 - From top to bottom: Original image; noisy image; denoised by NMF; denoised by DRNMF



- [11] Miao Zheng Xiaofei He Lijun Zhang, Zhengguang Chen, “Robust non-negative matrix factorization,” in *Front. Electr. Electron. Eng. China*, 2010.
- [12] Bin Shen, Luo Si, Rongrong Ji, and Baodi Liu, “Robust nonnegative matrix factorization via l_1 norm regularization,” in *arXiv:1204.2311*, 2012.
- [13] Deguang Kong, Chris Ding, and Heng Huang, “Robust nonnegative matrix factorization using l_{21} -norm,” in *Proceedings of the 20th ACM international conference on Information and knowledge management*, New York, NY, USA, 2011, CIKM '11, pp. 673–682, ACM.

Figure. 3. Denoise Result 3 - From top to bottom: Original image; noisy image; denoised by NMF; denoised by DRNMF

Electronic Health Card: Opportunities and Challenges

Hamid-Reza Firoozy-Najafabadi
Department of Computer Engineering
Science and Research Branch,
Islamic Azad University
Tabriz, Iran
hr.firoozy@iautash.ac.ir

Ahmad Habibzad Navin
Department of Computer Engineering
Science and Research Branch,
Islamic Azad University
Tabriz, Iran
habibi@iautash.ac.ir

Abstract— In recent years, information and communications technology (ICT) has been able to penetrate in the various fields of medical science and healthcare and has an important role in improving social health. In this regard the electronic health card (eHC) project has been introduced that is a replacement for insurance booklet and some countries have implemented eHC system. Electronic health card will have many benefits including, reducing the number of clerical and paper processes, costs, medical errors, more accurate follow-up of cases, integrity of patient information, eliminating the problem of illegible prescriptions, etc. Of course, this card also has challenges and obstacles. This paper first introduced the electronic health card and its benefits. Then the challenges of implementing this system will be examined and we will provide solutions for some of these challenges.

Keywords— *information and communications technology; electronic health card; Opportunities; challenges;*

I. INTRODUCTION

Increasing growth and development of ICT in recent years created dramatic changes in the social and economic life of human. And with creating appropriate contexts, it has given opportunities and new environment to economic and social activists to expand their activities and to provide them much better [1, 4]. Using information and communications technology have numerous advantages that some of them are: saving time, improving productivity and performance, reducing costs, providing better services to citizens, creating new job opportunities, economic growth, new management methods, development of international business, etc. Growth and development of ICT in various fields such as healthcare, business, banking, tourism, education, government and industry in different countries is increasingly pervasive. In the present situations, governments without the using ICT can't fulfill their social needs. And their position among the people and even international communities will lose [1].

One of the most important parts of ICT is designing and implementation of electronic service tools based on the society and citizen needs. One of the existing tools to provide electronic services is the smart cards. Depending on the application of these cards they are designed with various architectures. Smart cards using the new technology provide possibility of store, retrieve, maintenance, processing and information management. Today the process of issuing smart

cards for various applications is expanding and this subject will lead the society to the situation that each organization can issue one or more smart cards for their services [2, 5].

In this regard, insurance organizations in different countries have proceeded to create electronic health card (eHC) and replace it instead of paper insurance. This means that each person will have an electronic health card instead of paper insurance booklet. Therefore all medical and healthcare operations will be performed via this card. eHC will have many benefits such as reducing the number of clerical (paper) processes, costs, medical errors, more accurate follow-up of disease cases, integrity of information, etc [10, 9, 6]. Rather than advantages and applications mentioned to eHC card, this system has also some challenges that in this paper we will discuss about these challenges.

After this introduction, paper is organized as follows: Section 2 offers a general survey on smart cards and electronic health. In Section 3 electronic health card, benefits and challenges are discussed. In section 4 opportunities and challenges of eHC system will be discussed. Finally conclusion of this study is presented.

II. BACKGROUND

In most countries, healthcare services are offered via insurance booklet in healthcare centers. But time passing has shown that this method has disadvantages and problems that are [6, 8, 9, 10]:

- Time waste in performing the process
- Paper prescription illegibility
- Abuse of other people's booklets and cheating
- Renewal (credit extend) problems of booklet
- Inspection and management problems
- Increased costs due to paper and office processes
- Repeated prescriptions and treatments due to non-availability of patient dossier
- Medical errors due to the lack of accuracy and integrity of information
- Lack of careful follow-up of disease in a long term treatment

- Not confidential information for patients and access to disease of people by unauthorized persons

Disadvantages and problems mentioned, cause that the healthcare system in different countries think about a solution for problems of insurance booklet and hence electronic health card plan was introduced as a replacement for paper insurance booklet. This project is currently implemented in some countries. Following we will introduce and evaluate electronic health card.

A. Smart Card

There are usually two types of smart cards [2]:

1) *Memory Cards*: These types of smart cards have only memory and as software disc can store information. Memory cards can only executed predetermined operation and have not the ability to do other operation (e.g. phone cards).

2) *Cards Equipped with Processor*: This type of smart cards in addition to memory, have a processing unit. Therefore information that they have on the memory can be processed, deleted, added or modified (e.g. fuel card).

Smart cards are equipped with a processor of a small computer that has a chip and this chip includes a central processing unit and some memory. Electricity needed to run the personal computer is provided by the card reader device when it is connected. Also like any computer, these cards have an operating system that when making card it is stored in the memory. It should be noted that a smart card equipped with processor can perform cryptographic operations within itself [17, 2].

B. Electronic Health

Electronic health is an emerging field of collisions between medical informatics, public health and business that will be promoted via the Internet and related technologies. In other words, electronic health is a team effort of healthcare specialists and government using ICT for healthcare of people. Availability, cost, quality and portability are important issues in the electronic health and healthcare [3, 15, 16].

C. Electronic Health Record

Electronic health record (eHR) includes the healthcare information of peoples that is stored electronically during their life. In fact, eHC covers all functions of a traditional record with better quality and speed. This record provides an integrated source of people health information. In the eHR cases such as length of treatment, the prescription drugs, experiments results, genetic information, medical images, etc are stored [3, 18, 19].

Electronic health record should be available electronically to authorized providers of healthcare or any other authorized person. The eHR should be available in every place and time in order to support quality promotion of healthcare services. [19].

III. ELECTRONIC HEALTH CARD

A. Types of Electronic Health Card

Electronic health card (eHC) is a plastic card similar types of smart cards that are issued to patients or medical professionals. Two types of health smart cards that have a lot of application in healthcare are [6, 11]:

1) *Patient Health Card (PHC)*: This card is kept with the patient and contains various information such as identity information, management information, medical information, drug prescriptions, etc. The international standard organization (ISO) is specified standard 21549 for the "Health informatics - information of patient health card." Figure 1 shows the general structure of a patient health card [12].

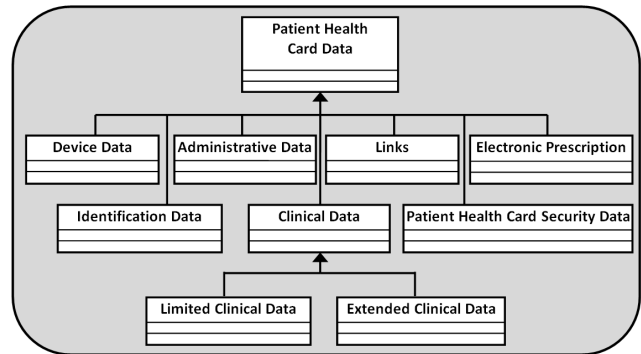


Figure 1. General patient health card structure

2) *Health Professional Card (HPC)*: This card is issued only for healthcare professionals. Healthcare professionals via this card and according to the access level that is defined by the position and specialty, will allow to read and write data on the patient health card. Indeed, the HPC is access key to the patient health card information. These cards includes identity information of healthcare specialists, personal code and their electronic signature.

B. Electronic Health Card System

Figure 3 shows a high level architecture of the eHC system. This system includes the following components [6, 10, 11, 14]:

- Health Card Issuing Organization
- Patient Health Cards
- Health Professional Cards
- Card Reader Devices
- Workstations (Computers)
- Centers Provider Health Services
- Pharmacies and Laboratories

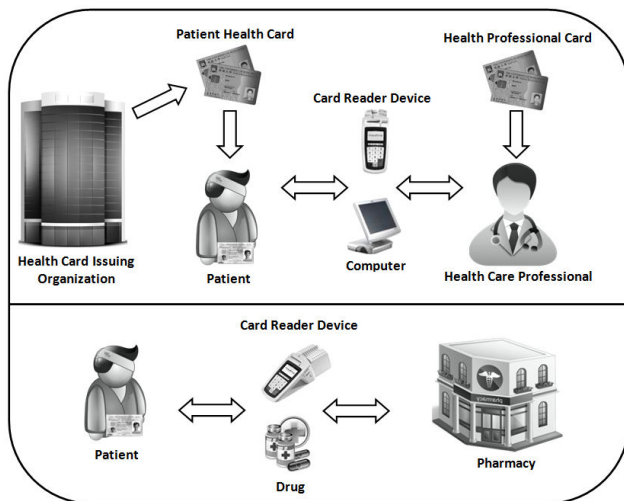


Figure 2. High level architecture of electronic health card system

C. Using Electronic Health Card System

In the healthcare centers (e.g. doctor office), there is a card reader device that connected to a computer. Doctor puts his/her card into card reader device and after entering of private code he/she can connect to the network.

At the time of referral to physician's office, patient delivers his/her card to reception part till its barcode will be scanned and patient name is added to the list of patients that will be visited by doctor. When patient enters doctor's room, puts his/her health card into card reader device and then registers his private code. Next, according to defined access level, doctor can observe different information of patient health card such as previous prescription, experiments, radiographs, etc. And also can register other information or create a new prescription on the PHC.

Finally, patient takes his/her PHC and if doctor prescribes a drug or an experiment, he/she will go to the pharmacy or laboratory. In those places, specialist put PHC into card reader device and according to defined access level, will have to needed information (e.g. drug prescription) [8, 10].

IV. OPPORTUNITIES AND CHALLENGES OF ELECTRONIC HEALTH CARD SYSTEM

A. Advantages of Electronic Health Card

Electronic health card system has special advantages as follows [6, 8, 9, 10, 15]:

- **Costs Reduction:** This project with regard to elimination of clerical (paper) processes, reducing costs related to issuing of insurance booklet, printing, extension, correction, etc.
- **Time Save:** with existence of eHC system, we don't need long time for completing forms and insurance booklets and these operations will be done electronically. For example the extension of insurance

credit will be done electronically very fast and information search accomplish with high speed.

- **Integration and Coordination of Information:** Electronic health card creates an integrated database of disease background. As a result it will keep information integration and decrease imbalance.
- **Troubleshooting Illegible Prescriptions:** Since all prescriptions are registered electronically on the eHC, it will eliminate problems related to illegible medical prescription.
- **Troubleshooting of Misusing from Insurance Booklets and Fraud:** After implementation of this system, insurance booklets are eliminated and patient authentication will be done electronically via eHC. So there aren't cases such as misuse from booklets of another individuals and also fraud.
- **More Accurate Follow-up of Disease Cases:** If there is an integrated network of patient information (e.g. electronic health record) and accurate records of their diseases, follow-up of treatment is done more easily and very.
- **Reducing Medicinal Errors:** As mentioned, with the implementation of this system, imbalance and conflicts will be resolved. Also medical errors and the economic losses will be reduced.
- **Better Management:** Due to availability of accurate statistics, planning and management in the fight against diseases and epidemics can be done better. Also the healthcare services are distributed fairly among citizens.
- **Ease of Inspection:** Financial-medicine inspection is done very fast and easily by agents of insurance organizations.
- **Increase of Security and Trust:** With regard to using new security mechanism in smart card for user authentication and defined access level to information. Unauthorized persons can't access to information of card. As a result, security and trust will be increased.

B. Challenges in Electronic Health Card

Rather than advantages and applications mentioned to electronic health cards, this card also has some challenges that should be managed accompanied with implementation. Some of these challenges include:

- **Authentication:** Authentication of patients and healthcare professionals is one of the most important challenges in the eHC system. The current method of user authentication is that the card and card reader device communicate with using APDU (Application Protocol Data Unit) packages and a common symmetric encryption key [17]. Using new technologies and techniques such as PKI (Public Key Infrastructure) or Biometric Techniques, it is possible to provide higher levels of trust and security for authentication [4]. PKI is a protocol that provided

asymmetrical encoding by public and private key pair. Biometric authentication techniques including fingerprint recognition, retinal scanning, hand geometry scanning and handwriting and voice recognition can be used too. These techniques are all based on the physical properties of a person [4].

- *Access control:* After user authentication in system, user access level to health card information and patient health records should be defined. In other words, all authorized individuals shouldn't access to entire information. But every person should use from available information according to defined access level. For example pharmacist can only use patient prescription therefore this plan need to issue different and appropriate permissions for system users [7].
- *Security and Privacy:* Most people don't like to share disease and personal information with others. Therefore, the electronic health card system should implement a special mechanism that ensuring privacy and security.
- *Security of Electronic Communications:* Communication and information transmission in the eHC system is based on electronic communications and this communication link may be attacked. Consequently the solutions for secure and reliable communication links should be available.
- *Availability:* The eHC system implementation only leads to full performance that this system be available in all healthcare centers (even in villages and small towns) including medical offices, clinics, laboratories, hospitals pharmacies, etc.
- *Citizens and Specialists Readiness:* One of the pre-requisite for eHC system is acceptability and tendency of citizens and healthcare professionals to use it. Because lack of people knowledge with this technology, may cause their resistance for using this system. So training and preparing people to use eHC system is necessary [4].
- *Low Band Width for Data Transfer:* Currently the network band width is very limited in many areas. So that fulfillment of needs of such a system would not be possible. Since all communications on this plan are electronic and are based on network communications. The solutions are necessary to improve communications and network bandwidth.
- *Lack of Integrated Information Networks:* The eHC system needs an integrated database of disease information for all peoples that has been available at any time and be accessible by authorized persons. Now there is no such database but by creating an electronic health record for each person, this problem can be solved.
- *Integrating The Electronic Health Record:* One of the main challenges in the eHC system is coordination and integration with electronic health record system. These two systems will be integrated and are able to

exchange information in future because the electronic health record is as a database for health electronic health card system. This requires that standards be considered properly during the implementation of both projects.

CONCLUSION

In recent years, information and communications technology (ICT) has been able to penetrate in the various fields of medical science and healthcare and has an important role in improving social health. In this regard the electronic health card (eHC) project has been introduced that is a replacement for insurance booklet. Electronic health card will have many benefits including, reducing the number of clerical and paper processes, costs, medical errors, more accurate follow-up of cases, integrity of patient information, eliminating the problem of illegible prescriptions, etc. In this paper we first introduced the electronic health card and its benefits. Then we discussed about the challenges of implementation this system and provided solutions for some of the challenges.

REFERENCES

- [1] P. Warren, J. Davies and David Brown, "ICT Futures: Delivering Pervasive, Real-time and Secure Services", John Wiley, 2008.
- [2] W. Rankl, W. Effing, "Smart Card Handbook", 4th ed., John Wiley, 2010.
- [3] R. Garte, "Electronic Health Records: Understanding and Using Computerized Medical Records", 2nd ed., Pearson Education, 2011.
- [4] H. R. Firoozy-Najafabadi and S. Pashazadeh, "Mobile Police in Mobile Government", 5th IEEE International Conference on Application of Information and Communication Thechnologies, pp. 118-122, 2011.
- [5] K. M. Sheller and J. D. Procaccino, "Smart card evolution", Magazine Communications, ACM, Vol. 45 No. 7, pp. 83-88, 2002.
- [6] N. Ernstmann, O. Ommen, M. Neumann, A. Hammer, R. Voltz and H. Pfaff, "Primary Care Physician's Attitude Towards the GERMAN e-Health Card Project-Determinants and Implications", Springer J Med Syst., pp. 181-188, 2008.
- [7] B. Blobel, "Authorisation and access control for electronic health record systems", International Journal of Medical Informatics, Vol. 73 No. 3, pp. 251-257, 2004.
- [8] M. Conrick, "Health Informatics: Transforming Healthcare with Technology", Thomas Nelson Australia, 2006.
- [9] D. Dieng, "The keys of success of a health card project: the lessons learned from 9 health cards project", Proc. of the 10th IEEE Symposium on Computer-Based Medical Systems, pp. 227-230, 1997.
- [10] Ch. T. Liua, P. T. Yanga, Y. T. Yeha and B. L. Wang, "The impacts of smart cards on hospital information systems—An investigation of the first phase of the national health insurance smart card project in Taiwan", International Journal of Medical Informatics, Vol. 75, pp. 173-181, 2006.
- [11] P. Pharow and B. Blobel, "Security Infrastructure Requirements for Electronic Health Cards Communication", Connecting Medical Informatics and Bio-Informatics, pp. 403-408, 2005.
- [12] ISO standard 21549, "Health Informatics – Patient health card data", 2003.
- [13] G. Yee, L. Korba and R. Song, "Ensuring Privacy for E-Health Services", Proc. of thhe 1st IEEE International Conference on Availability, Reliability and Security, pp. 8, 2006.
- [14] M. Zwicker, J. Seitz and N. Wickramasinghe, "An Approach of a Telematics Infrastructure for the German Electronic Health Card", Proc. of the 23rd IEEE Symposium on Computer-Based Medical Systems, pp. 438-444, 2010.

- [15] M. Torabi and R. Safdari, "Electronic Health", Secretariat of Iranian Informing Council, 2006.
- [16] B. Fathi, H. Riazi and E. Bitaraf, "Electronic Health: Components and Standards", Iranian Information Technology Council, Electronic Health Workgroup, 2007.
- [17] A. Safaei, "Security and Privacy of Smart Cards", Iranian Network Journal, 2007.
- [18] M. Torabi and R. Safdari, "Electronic Health", Jafari, 2009.
- [19] M. Ahmadi, P. Rezaei and L. Shahmoradi, "Electronic Health Record: Structure, Content and Evaluation", Jafari, 2008.

Application of Functional State Modelling Approach for Yeast *Saccharomyces cerevisiae* Fed-batch Fermentation Modelling

SANDIS VILUMS¹, EMILS KOZLINSKIS², VALTERS BRUSBARDIS¹

¹Biosystems Group, Faculty of Information Technology, Latvia University of Agriculture
Lielā iela 2, LV-3001 Jelgava, Latvia

²Faculty of Food Technology, Latvia University of Agriculture
Lielā iela 2, LV-3001 Jelgava, Latvia
sandis.vilums@gmail.com

Abstract—Continuous estimation of bioprocess state during yeast fed-batch fermentation requires using accurate mathematical model and parameters that can describe the actual state of the yeast growth during fermentation. Application of functional state modelling approach for the yeast *Saccharomyces cerevisiae* fed-batch fermentations is presented. The main advantage of the modelling approach over global model with complex structure is that the parameters of each local model can be estimated separately from the other local model parameters. Moreover, this approach makes it easier to perform model simulation and parameter estimation compared with the complex global model. Obtained experimental data from yeast fed-batch fermentations show sufficiently good match with the curves simulated using functional state fermentation model.

Keywords: *Bioprocess modelling, fed-batch fermentation, parameter estimation.*

I. INTRODUCTION

The yeast *Saccharomyces cerevisiae* is one of the most relevant microorganisms in biotechnology industry. In view of increasing importance of ethanol as an alternative source for chemicals and liquid fuel a great deal of research interest in ethanol fermentation has been generated [1–3].

Continuous evaluation of process parameters during the cultivation is essential for obtaining mathematical model of fermentation that can simulate the process. Global process models are generally used for bioprocess modelling [4], [5]. The main disadvantage of such approach is complexity of model structure and a large number of model parameters, which complicates model simulation and parameter estimation.

The functional state modelling approach is an alternative concept of bioprocess modelling. In functional state the process is described by local model that is valid in actual state only. A set of local model together with functional state equations can be used to describe, monitor and control the yeast growth process. Several authors have already presented benefits of this approach for more accurate and detailed process modelling [6–11].

This article aims to demonstrate the functional modelling approach benefits of yeast fed-batch fermentation process modelling. In case of high coincidence of model and experimental data this approach can be used not only for process predictions but also for early stabilization of process using fermenter control program.

II. FED-BATCH FERMENTATION

The organism used in this study was yeast *Saccharomyces cerevisiae* DY 7221. 2 liters of batch medium was prepared according to the requirements of *S. cerevisiae*, containing glucose 5.0 gL⁻¹, (NH₄)₂HPO₄ 0.85 gL⁻¹, K₂HPO₄ 0.5 gL⁻¹, MgSO₄·7H₂O 0.1 gL⁻¹, ZnSO₄·7H₂O 0.00025gL⁻¹, FeCl₃·6H₂O 0.00017 gL⁻¹, MnSO₄·5H₂O 0.00025 gL⁻¹, yeast extract 5 gL⁻¹. 2 liters of feeding medium contained glucose 166 gL⁻¹, (NH₄)₂HPO₄ 1.2 gL⁻¹, ZnSO₄·7H₂O 0.001 gL⁻¹, FeCl₃·6H₂O 0.00017 gL⁻¹, MnSO₄·5H₂O 0.001 gL⁻¹, (NH₄)₂SO₄ 14.5 gL⁻¹, yeast extract 30 gL⁻¹. Batch and feeding medium was sterilized in autoclave for 30 min at 110 °C. Dry pellets of *S. cerevisiae* were used for inoculation in yeast fermentations directly to batch medium with initial biomass concentrations of ~4 g/L. Initial OD of fermentation was 8.3 at 600nm.

The fed-batch fermentations were carried out in a stirred tank fermenter (Sartorius Stedim Biostat Bplus) with a working volume of 5 liters. Temperature was controlled at 30.0 ± 0.2 °C, pH level at 4.3 ± 0.3, dissolved oxygen partial pressure at 40 ± 3 %. Agitation and dissolved oxygen partial pressure value were ensured with two “rushton” style impellers varying from 200 to 800 rpm. Airflow was ensured to 1.7slpm. 5 ml of sample were taken every 1.5 hours for the entire fermentation cycle, which was terminated after 21 hours. Yeast biomass growth was evaluated by spectrophotometric measurements at 600 nm in a *Jenway 6300* UV-visible spectrophotometer. The concentration of glucose was determined used glucose measurement device *Accu-Chek Active*. The concentration of ethanol in the medium was determined by liquid chromatography (HPLC).

III. FERMENTATION MODELLING

Dynamic mass balances are the traditional chemical engineering approach to state estimation in bioreactors.

Approach uses dynamic balances at the reactor scale and reasonable assumption regarding the regulatory structure of the organism. Making accurate measurements of bioreactor process variables, the possibility of using mass balance models has a high chance of success [12].

The rates of cell growth, glucose consumption, ethanol production, oxygen concentration and volume are described for all functional states with mass balance equations as follows:

$$\frac{dX}{dt} = \mu X - \frac{F}{V} X \quad (1)$$

$$\frac{dS}{dt} = -q_S X + \frac{F}{V} (S_{in} - S) \quad (2)$$

$$\frac{dE}{dt} = q_E X - \frac{F}{V} E \quad (3)$$

$$\frac{dO}{dt} = -q_O X + k_L a (O^* - O) \quad (4)$$

$$\frac{dV}{dt} = F + F_A + F_B + F_{ANT} - F_{SMP} - F_E - F_C \quad (5)$$

where X - concentration of biomass, $g \cdot L^{-1}$;
 μ - specific cell growth rate, h^{-1} ;
 S - concentration of substrate (glucose), $g \cdot L^{-1}$;
 S_{in} - concentration of feeding substrate (glucose), $g \cdot L^{-1}$;
 q_S - specific substrate consumption rate, h^{-1} ;
 E - ethanol concentration, $g \cdot L^{-1}$;
 q_E - specific ethanol production rate, h^{-1} ;
 O - oxygen solubility, $g \cdot L^{-1}$;
 q_O - specific oxygen consumption rate, h^{-1} ;
 $k_L a$ - volumetric oxygen transfer coefficient, h^{-1} ;
 O^* - maximal solubility of oxygen, $g \cdot L^{-1}$;
 V - volume, L;
 F - feeding rate, $L \cdot h^{-1}$;
 F_A - acid consumption rate, $L \cdot h^{-1}$;
 F_B - base consumption rate, $L \cdot h^{-1}$;
 F_{ANT} - antifoam consumption rate, $L \cdot h^{-1}$;
 F_{SMP} - sampling rate, $L \cdot h^{-1}$;
 F_E - evaporation rate, $L \cdot h^{-1}$];

F_C - carbon loss rate, $L \cdot h^{-1}$.

A substrate such as glucose is consumed by yeast to produce a number of carbon intermediates as well as to provide energy. The yeast then utilizes the carbon intermediates to synthesize a new cell material. If the sugar concentration in the medium in an aerobic yeast growth process exceeds a certain level, called the critical sugar mass concentration (S_{crit}), a part of the sugar is metabolized in ethanol. In the case of fed-batch cultivation S_{crit} is assumed to be 0.05 g/L. A critical level of dissolved oxygen concentration for yeast growth process is assumed to be 18%. The whole yeast growth process can be divided into at least five functional states in fed-batch cultures [6], [9].

In the case of fed-batch cultivation six phases are identified (Fig. 1.):

1. The first functional state (I) is called *the first ethanol production state*. The process is defined to be in this state when the sugar mass concentration is above the critical level and there is sufficient dissolved oxygen. Ethanol is produced in this state.
2. The second functional state (II) is the *mixed oxidative state*. The process enters this state when the sugar concentration decreases to be equal to or below the critical level and there is sufficient dissolved oxygen in the medium. The process remains in this state as long as these conditions are met. Both sugar and produced ethanol are co-metabolised through the oxidative pathways in the state.
3. The third functional state (III) is the *complete sugar oxidative state*. The process is defined to be in this state when there is no ethanol available, the sugar concentration is not higher than the critical level and the dissolved oxygen is above its critical level (O_{crit}). Sugar is completely oxidised to water and carbon dioxide in this state.
4. The fourth functional state (IV) is called the *ethanol consumption state*. The process is defined to be in this state when ethanol is available but there is no sugar in the medium, and the dissolved oxygen concentration is above the critical level. Ethanol is the only carbon source for yeast growth.
5. The fifth functional state (V) is the *second ethanol production state*. The conditions for this state are that both concentrations, for sugar and dissolved oxygen, are below the corresponding critical levels. When the dissolved oxygen becomes the limiting factor for yeast growth, ethanol is produced.
6. The sixth functional state (VI) is the *dissolved oxygen limitation state*. The process is defined to be in this state when the sugar concentration is above the sugar critical level and dissolved oxygen is below the critical level so it becomes the limiting factor for the yeast growth.

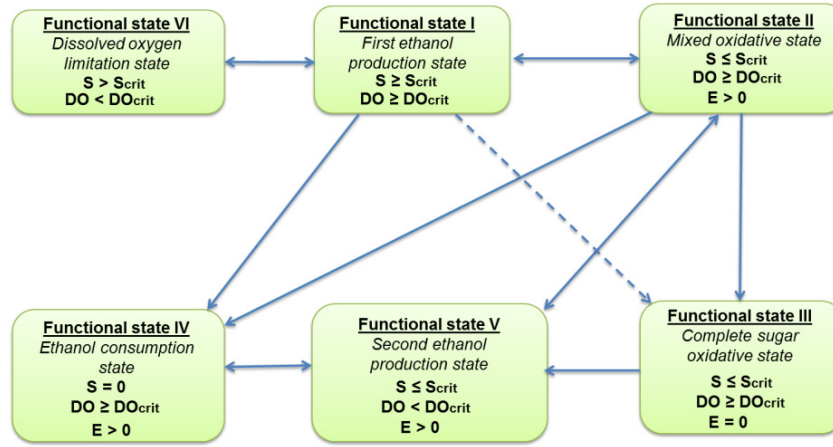


Fig. 1. Identified functional states of the yeast fed-batch cultivation

[9] assume that in principal, the functional state (I) can appear in all batch, fed-batch, and continuous yeast growth processes. A yeast growth process switches from one functional state to another like a state machine or automation familiar in computer science. Parameter functions of the local models in the states (I), (II), (III), (IV), (V) and (VI) are presented in Table 1.

After entering state II yeast cell begin to synthesize the enzymes for gluconeogenesis so that cells can utilize ethanol as the carbon-source for growth. This causes a lag in the yeast growth. The lag term for functional state can be calculated by equation (6).

Table 1

Parameter functions of the local models[6]

Parameter functions	State I	State II	State III	State IV	State V	State VI
μ	μ_{\max}	$\mu_{2S} \frac{S}{S+k_S} + \mu_{2E} \frac{E}{E+k_E}$	$\mu_3 \frac{S}{S+k_S}$	$\mu_{2E} \frac{E}{E+k_E} \eta$	$\mu_{3S} \frac{S}{S+k_S} + \frac{O_2}{O_2+k_{O_2}}$	$\mu_6 \frac{S}{S+k_S}$
q_S	$\mu_{\max} \frac{S}{S+k_S} Y_{X/S}$	$\mu_{2S} \frac{S}{S+k_S} Y_{X/S}$	$\mu_3 \frac{S}{S+k_S} Y_{X/S}$	0	$\mu_3 \frac{S}{S+k_S} Y_{X/S}$	$\mu_6 \frac{S}{S+k_S} Y_{X/S}$
q_E	$(q_S - q_{Scrit}) Y_{E/S}$	$-\frac{\mu_{2E}}{Y_{E/X}} \frac{E}{E+k_E}$	0	$-\mu_{2E} \frac{E}{E+k_E} Y_{E/X} \eta$	$\mu_3 \frac{S}{S+k_S} Y_{X/S} * \frac{k_{O_2}}{O_2+k_{O_2}} Y_{S/E}$	$\frac{\mu_6}{Y_{O/X}} \frac{O_2}{O_2+k_{O_2}}$
q_O	$\frac{\mu_{\max}}{Y_{O/S}}$	$q_E Y_{O/E} + q_S Y_{O/S}$	$q_O \frac{S}{S+k_S}$	$\mu_{2E} \frac{E}{E+k_E} Y_{O/E} \eta$	$q_{O_2} \frac{S}{S+k_S} * \frac{O_2}{O_2+k_{O_2}}$	$\frac{\mu_6}{Y_{O/X}} \frac{O_2}{O_2+k_{O_2}}$

where μ_i – maximum specific growth rates, h^{-1} ;
 k_S, k_E, k_O – saturation constants, $g \cdot L^{-1}$;
 $Y_{XS}, Y_{ES}, Y_{OS}, Y_{EX}, Y_{OX}, Y_{OE}$ – yield coefficients, $g \cdot g^{-1}$;
 q_S – specific substrate consumption rate, h^{-1} ;
 q_{Scrit} – critical specific substrate consumption rate, h^{-1} ;
 q_E – specific ethanol consumption rate, h^{-1} ;
 q_O – specific oxygen consumption rate, h^{-1} ;
 η – lag term, h.

$$\eta = 1 - \exp\left(-\frac{t-t_m}{t_1}\right) \quad (6)$$

where t – the current process time, h;

t_m – time point of involving in lag phase, h;
 t_1 – the length of lag phase, h.

IV. RESULTS AND DISCUSSION

The functional state model of yeast fed-batch process was developed in MATLAB environment. The model consists of script M-Files and experimental data CSV files. Experimental data obtained from two similar fed-batch fermentation of yeast *Saccharomyces cerevisiae* DY7221. The experimental data contains on-line measurements of pO₂, stirrer rpm, feedrate and airflow. Off-line data include measurements of biomass, substrate (glucose) and ethanol. The fed-batch

experiment measured and model simulated curves are presented in the Fig 2.

The process was carried out in 21 hours. Model simulated curves X, S, E show good accordance with measured experimental data. The actual functional state is identified automatically from actual substrate and dissolved oxygen saturation concentration. The functional state plot (Fig. 2) shows that switching from the 1st to another functional state didn't occur because process substrate and dissolved oxygen concentration was above critical values all through the process. From the beginning of process till 8th hour 1st functional state with specific growth rate 0.125 h⁻¹ was identified. After 8th hour specific growth rate of biomass decreased to 0.056 h⁻¹. The estimated model parameters are presented in Table 2.

Table 2

Fed-batch experiment estimated values of functional state parameters

Parameters of State I	Estimated value
$\mu_{\max 1}$	0.125
$\mu_{\max 2}$	0.056
K_S	0.173
Y_{XS}	0.126
q_{Scrit}	0
Y_{ES}	0.394
Y_{OS}	0.2

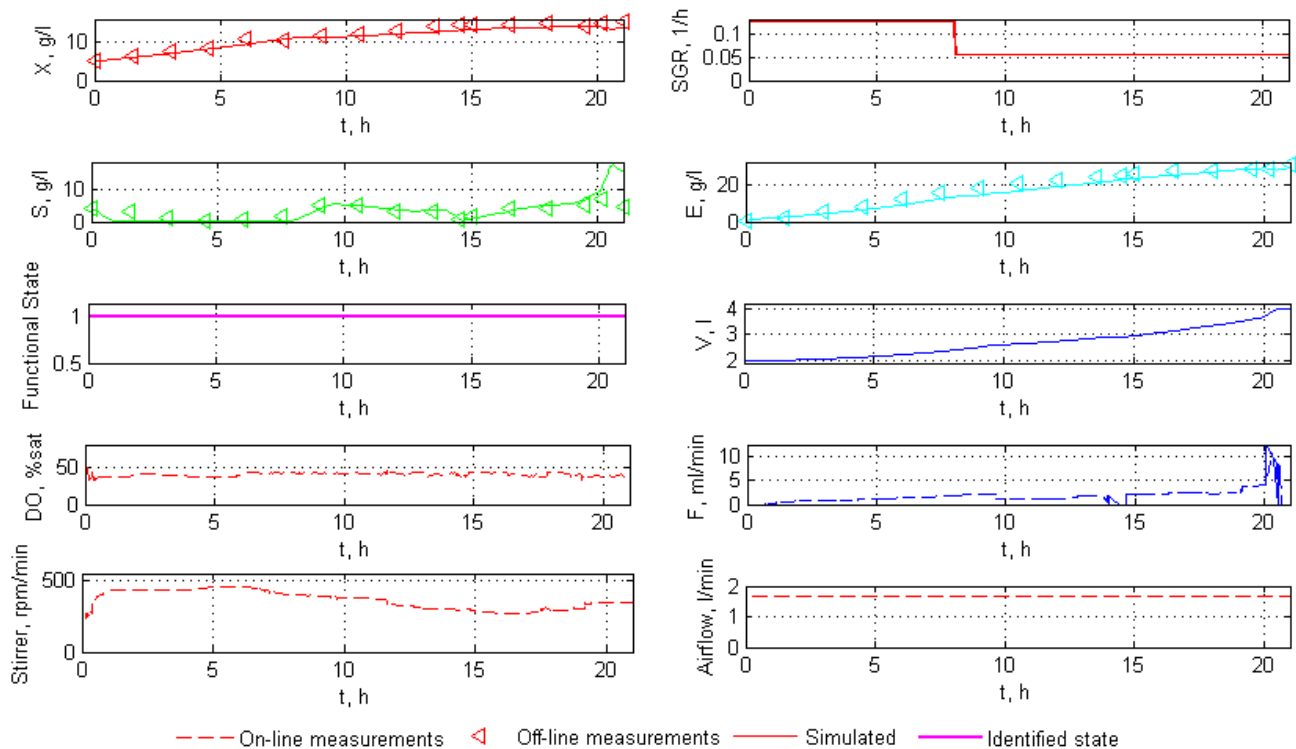


Fig. 2. Fed-batch fermentation measured, model simulated and identified state curves

V. CONCLUSION

The application of the functional state modelling approach for yeast fed-batch fermentation was presented in this paper. This approach shows substantial benefits for using it as foundation for the yeast fed-batch fermentation model based control system development. The yeast growth functional state estimation is valuable for clarifying appropriate parameters that can adequately describe the actual physiological state of yeast growth process. Model simulated curves for X (biomass), S (substrate) and E (ethanol) showed good accordance with measured

experimental and model simulated data for both fed-batch fermentations. The functional state modelling approach parameters for actual functional state were automatically estimated using Simulink model and embedded Matlab code optimization procedure by varying parameter value to reach the best experimental data and model simulated curve consistency. Nelder-Mead optimization method was used for parameter estimation. Improvements in yeast fed-batch fermentation model with automatic parameter estimation for all identified functional states is planned to realize in future.

ACKNOWLEDGEMENT

This paper was written by financial support of European Structural Fund – Project “Establishment of Latvia interdisciplinary interuniversity scientific group of systems biology” – realized by Latvia University of Agriculture (contract no. 2009/0207/1DP/1.1.1.2.0/09/IPIA/VIAA/128).

REFERENCES

- [1] A. Kumoro, G. Ngoh, M. Hasan, F. Chew, and M. Tham, “Production of Ethanol by Fed-Batch Fermentation,” *Pertanika Journal of Science and Technology*, vol. 17, no. 2, pp. 399–408, 2009.
- [2] W.-H. Hunag, G. S. Shieh, and F.-S. Wang, “Optimization of fed-batch fermentation using mixture of sugars to produce ethanol,” *Journal of the Taiwan Institute of Chemical Engineers*, Sep. 2011.
- [3] J. Hjersted and M. A. Henson, “Population Modeling for Ethanol Productivity Optimization in Fed-Batch Yeast Fermenters,” *Cell Cycle*, pp. 3253-3258, 2005.
- [4] F. Renard, a Wouwer, S. Valentinotti, and D. Dumur, “A practical robust control scheme for yeast fed-batch cultures – An experimental validation,” *Journal of Process Control*, vol. 16, no. 8, pp. 855-864, Sep. 2006.
- [5] Z. Nagy, “Model based control of a yeast fermentation bioreactor using optimally designed artificial neural networks,” *Chemical Engineering Journal*, vol. 127, no. 1–3, pp. 95-109, Mar. 2007.
- [6] T. Pencheva, I. Hristozov, S. Tzonkov, and B. Hitzmann, “Functional State Modelling of *Saccharomyces cerevisiae* Cultivations,” *Yeast*, vol. 1, pp. 1-15, 2004.
- [7] I. Hristozov, T. Pencheva, S. Tzonkov, and B. Hitzmann, “Functional State Modelling Approach for Batch Cultivation of *Saccharomyces cerevisiae*,” *Biomedical Engineering*, vol. 19, no. 1, pp. 69-74, 2005.
- [8] T. Pencheva and I. Hristozov, “Modelling of Functional States during *Saccharomyces cerevisiae* Fed-batch Cultivation,” pp. 8-16, 2005.
- [9] O. Roeva, T. Pencheva, U. Viesturs, and S. Tzonkov, “Modelling of Fermentation Processes Based on State Decomposition,” pp. 1 - 12, 2006.
- [10] O. Roeva et al., “Multiple model approach to modelling of *Escherichia coli* fed-batch cultivation extracellular production of bacterial phytase,” *Electronic Journal of Biotechnology*, vol. 10, no. 4, Oct. 2007.
- [11] S. Vilums and O. Grigs, “Application of functional state modelling approach for yeast *Saccharomyces cerevisiae* batch fermentation state estimation,” in *5th International Scientific Conference “Applied Information and Communication Technologies,”* 2012, pp. 300-305.
- [12] C. Komives and R. S. Parker, “Bioreactor state estimation and control,” *Current Opinion in Biotechnology*, vol. 14, no. 5, pp. 468-474, Oct. 2003.

FETAL ECG REPRESENTATION USING RECURRENCE PLOT ANALYSIS

¹Elif Tuba Celik, ¹Bogdan Hurezeanu, ²Angela Digulescu, ²Madalina Mazilu

¹Polytechnic University of Bucharest, Romania

²Military Technical Academy, Bucharest, Romania

Abstract – The field of electrocardiography (ECG) has been in existence for over a century, but the analysis of fetal ECGs is still in its infancy. This is, partly due to a lack of availability of elegant standard databases, partly due to the relatively low signal to noise ratio of the fetal ECG compared to maternal ECG.

In this paper we applied recurrence plot analysis method for noninvasively recorded abdominal ECG. This method is based on the analysis of recurrences of the state space trajectory of the system under analysis that has been employed especially for studying chaotic dynamical systems like ECG signals. The method is successfully applied to reduce the effect of noise on the ECG signal and represent fetal component that is not completely hidden by the maternal ECG.

Keywords – ECG signals, recurrence plot analysis, denoising, state space representation

I. INTRODUCTION

There are a lot of stress elements on the fetus during labour and it is highly recommended a carefully follow its evolution. Doppler echography is widely used to obtain qualitative information on the fetal status. However, one of the main causes of natal illness and death is hypoxia. Fetal ECG (fECG) presents within its morphological and temporal characteristics, essential information to identify hypoxia, its analysis has become an important element in studying fetal health during pregnancy.

fECG analysis as important as it is, the more difficult is to extract considering the presence of the mother ECG (mECG) and other noise sources. Abdominal ECG is illustrated in Fig.1.

Different techniques for noise removal or/and detection of fetal waveforms have been used. The most recent ones include algorithms that are based on singular value decomposition [1], auto and cross-correlation techniques [2], adaptive filtering [3], orthogonal basis functions [4], IIR adaptive filtering combined with genetic algorithms [5], even Kalman filtering which is expected to improve the denoising and extraction of fECG components, fuzzy logic, frequency tracking and real-time signal processing. In addition, independent component analysis for blind source separation has also been applied [6].

The wavelet transform (WT) is another approach that has been proposed for fetal ECGs processing. Also, the Gabor-8 power wavelets combined with the application of the Lipschitz exponents were applied to extract fECG[7]. Furthermore, wavelet-based multiresolution analysis has been proposed for noise removal.

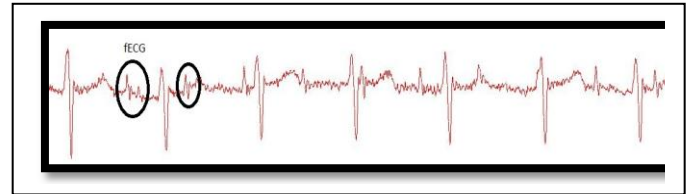


Figure 1. Abdominal ECG recording, Fetal and maternal influence.

As stressed in [8] the fECG extraction poses problems which can't be easily solved by conventional filtering technique so in this paper we propose a different method of filtering namely the analysis of recurrences in the state space. The method proposed is the outcome of nonlinear dynamical systems analysis and has its number of advocates also, is of general interest outside the domain of fECG.

The state-space representation has the benefit of allowing the study of the evolution of the signal dynamics using state-space approaches, furthermore the visual representation of these dynamics and image analysis to determine the morphology of a system associated within the image.

However, there are many issues such as the interpretation, stability, robustness, and noise-sensitivity of the extracted components. These issues are left as open problems and require further studies by using realistic models of these signals.

To conclude, we present the organization of paper. Section II presents a short presentation of the used method to analyze time series by using the trajectory's recurrences in the state space and plotting its recurrent states. Section III shows a method to reduce noise directly from the recurrence plot, by filtering it. In section IV we present the results obtained by applying the method described previously. The paper is closed by some conclusions.

II. RECURRENCE PLOT ANALYSIS

Most often we might have only a single sequence of measurements to identify the state of the system observed at successive moments (a time series) [9]. Our aim in this study is characterization of time series recorded from the abdominal ECG and their corresponding two-dimensional images - Recurrence Plots.

Recurrence Plots (RPs) are relatively new technique for the qualitative assessment of time series [10]. With RP, one can graphically detect hidden patterns and structural changes in data or see similarities in patterns across the time series under study. The fundamental assumption underlying the idea of the recurrence plots is that an observable time series is the realization of some dynamical process, the

interaction of the relevant variables over the time. Because the effect of all the other (unobserved) variables is already reflected in the series of the observed input, one can recreate a topologically equivalent picture of the original multidimensional system behavior by using the time series of a single observable variable. The method developed by Eckmann give the possibility to represent a m-dimensional state space trajectory in a two dimensional representation.

Recurrence Plot Analysis can be summarized as follows:

- Given a series of observations $\{s_1, s_2, \dots, s_N\}$, the first step is to reconstruct the state space of the time series. This reconstruction is done by *time-delay embedding* which uses two parameters: m – the dimension of the reconstructed state space and d – the time delay between two successive elements of the reconstructed state vector. For choosing d the average mutual information method is used and for m the false nearest neighbors method.
- The next step is the transition from the reconstructed state space trajectory to the recurrence matrix. At this step another two parameters are necessary: the metric used and ϵ – the size of the neighborhood. The first parameter doesn't have a strong effect on the resulting recurrence matrix. It is often used the L_∞ metric, for its high computing speed. On the other hand ϵ is not that easy to choose. Marwan [10] suggests it should be five times greater than the variance of the observational noise. (This way the effect of a low level noise added to the time series is reduced.)
- The third step of the method consists in the interpretation of the resulting recurrence matrix. This interpretation can be done visually by experienced ones, but in order to automatize this activity, a method to **quantify the recurrences**, called Recurrence Quantification Analysis (RQA), is used.

Three kinds of RQA measures are used:

- measures based on recurrence points density – RR (Recurrence Rate);
- measures based on the quantification of diagonal lines parallel to the recurrence plot's LOI (Line of Identity – the main diagonal of the recurrence matrix) – as is, for example, DET (Determinism);
- measures based on the quantification of the vertical lines from the recurrence plot – as is, for example, LAM (Laminarity).

Recurrence quantification analysis can be done on the entire recurrence matrix, or on small windows from it, taken along the LOI. This second version gives the opportunity to analyze the time variation of the RQA measures, and it needs two parameters:

- w – the size of the analysis window;
- ws – the shift of the analysis window.

The values of these parameters can strongly affect the computed RQA measures.

The main advantage of the recurrence plots over another widely used techniques as for example Fourier analysis, is that they preserve both temporal and spatial dependence in the time series. Even though Fourier analysis reveals the distribution of spectral frequencies, it doesn't show how self-similar resonant frequencies are patterned as a function of time. Yet, RP is mostly a qualitative technique and the precise meaning of the patterns is unknown.

III. THE EFFECT OF NOISE ON THE ECG SIGNAL

Biomedical signal means a collective electrical signal acquired from any organ that represents a physical variable of interest where the signal is considered in general a function of time and is describable in terms of its amplitude, frequency and phase. FECG is a biomedical signal that gives electrical representation of fetal heart rate to obtain the vital information about the condition of the fetus during pregnancy and labor from the recordings on the mother's body surface. The FECG signal is a comparatively weak signal (less than 20 percent of the mother ECG) and often embedded in noise. The fetal heart rate lies in the range from 1.3 Hz to 3.5 Hz and sometimes it is possible for the mother and some of the fetal ECG signals to be closely overlapping. The FECG is very much related to the adult ECG shown in Fig. 2, containing the same basic waveforms including the P-wave, the QRS complex, and the T-wave.

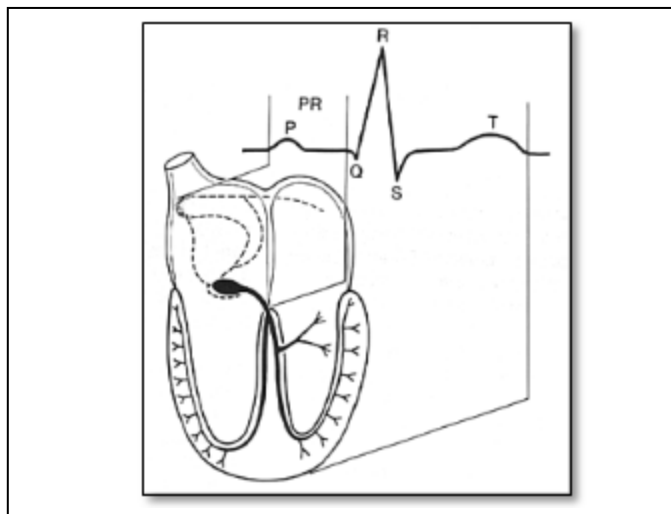


Figure 2. Anatomy of the heart with assignment of P, Q, R, S, T, and P waves

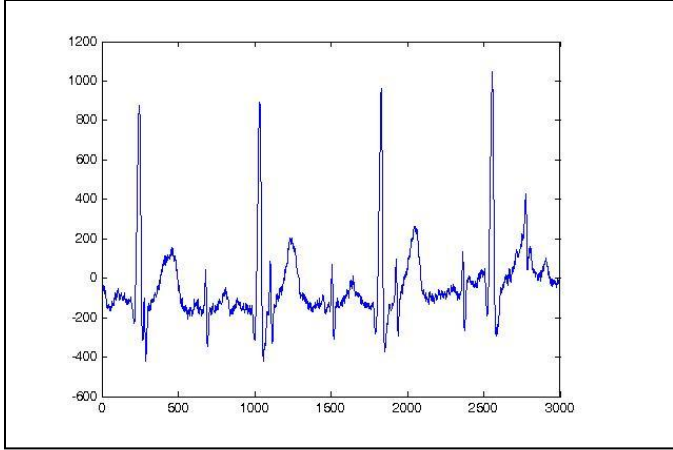


Figure 3. Signal representation of noisy abdominal ECG signal

Recent studies have shown that abdominal ECG signal is composed of mECG and fECG signals as discussed earlier in this paper. Because the study of abdominal signals is non-invasive method, this type of analysis has become important in order to obtain fECG. The fetal ECG contains potentially valuable information that could assist clinicians in making more appropriate and timely decisions during labour, but the fECG signal is vulnerable to noise and difficulty of processing it accurately without significant distortion has impeded its use. A number of difficulties and complication are associated with recording the abdominal ECG. Electrical activity recorded in Fig. 3 from the maternal abdomen suppresses the fECG (the magnitude of the fECG signal at the maternal abdomen is of the order of several micro volts), which is a fraction of the mECG amplitude recorded at the maternal abdomen. The abdominal ECG contains a weak fetal ECG signal, a relatively sound maternal ECG, maternal muscle noise (electromyographic activity in the muscles of the abdomen and uterus) and respiration, mains coupling, and thermal noise from the electronic equipment (electrodes, amplifiers, etc.), power line interference (A/C) and Baseline Wandering [11].

The question is whether in the presence of noise in the time series recurrence plot analysis applied to it can still offer valid results. Further we study the effect of noise on the recurrence plot.

IV. RESULTS

Our goal here is to obtain, from the noisy ECG signal, amplitude variations of the transients (both of mECG and fECG) which enables us to estimate the positions in time as well as the durations of the transients that are present in the analyzed signal.

$$DM, RM \in \mathbb{M} \quad M, M (\square)$$

$$DM_{i,j} = \|\vec{r}_i - \vec{r}_j\|, RM_{i,j} = H(\varepsilon - DM_{i,j}). \quad (1)$$

$$i, j = \overline{1, 2, \dots, M}, H - \text{Heaviside function}$$

Computing distance matrix and recurrence matrix, we obtain a recurrence plot as visual. The choice of the threshold to use for the binarization of the distance matrix is very important, as the interpretability of the obtained recurrence matrix depends on it [12]. This parameter must also reduce as much as possible the noise sensitivity in the original time series. We have chosen this parameter as 0.5, the average distance between successive points of the trajectory (the r_i, r_j vectors in Equation (1)).

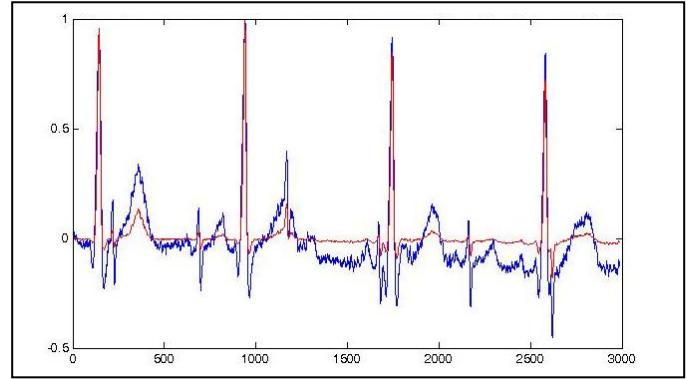


Figure 4. Signal representation of filtered matrix with the threshold 0.5: original signal (blue), filtered signal (red).

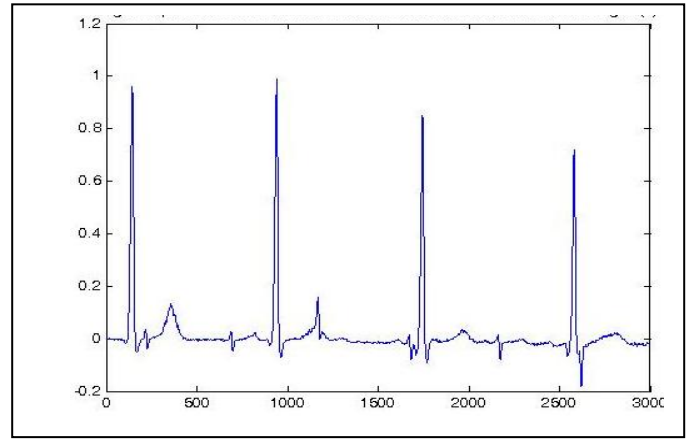


Figure 5. Signal representation of filtered matrix with the threshold 0.5: filtered signal (b).

The filtered signal is obtained from the convolution between the original signal and the sum of the columns of the distance matrix as shown in Fig. 4. The sum of the columns can be perceived as a 'mask' which shows the transitory states and uniforms the noise. The 'mask' over the original signal succeeds to highlight the mother ECG, as well as the fetal ECG.

At a closer look, this convolution allows us to detect more easily the points R peak of the mother together with the fetal too. Moreover, in Fig. 5 the overlapping ECGs between the mother and the child is more obvious in the case of the filtered signal. Another major advantage of these filtering techniques represents the fact that this method does not affect

the spectrum, the spectrum of the filtered signal being approximately the same in Fig.6.

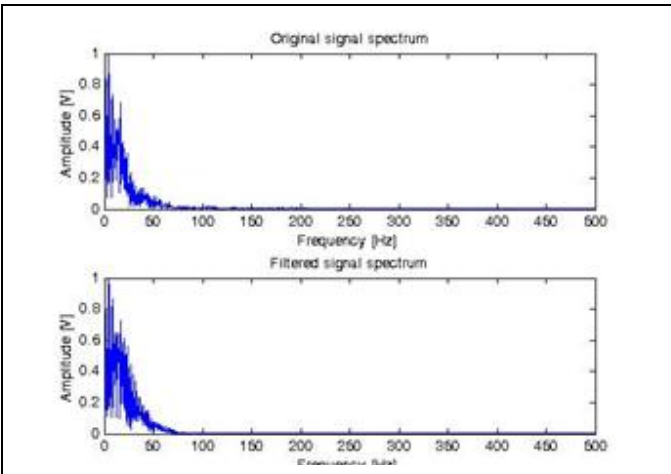


Figure 6. Original and filtered signal spectrum

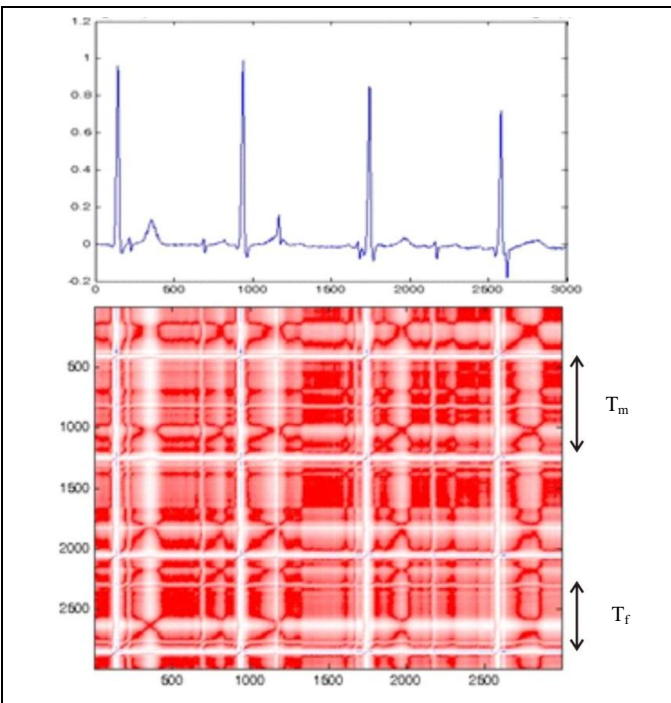


Figure 7. . Filtered distance matrix with 0.5 threshold

Based on the distance matrix there can be seen the transitions in Fig. 7: white bands, white vertical/horizontal lines all along the entire matrix. This means that all the points contained in that region have a totally different behavior from the others, being placed in a whole new area in the phase space, far enough from the all others points. The points that are close in the phase space are represented in red. In this case, these points represent generally the noise. One of the parameter of signal is periods of maternal and fetal are represented (T_m , T_f).

CONCLUSIONS

In our paper we applied the recurrence plot analysis method to abdominal ECG measurements in order to obtain the main parameters of maternal and fetal ECG. Our first goal was to obtain a denoising abdominal ECG signal from the recurrence plot representation in order to detect more easily the main parameters of the mECG and fECG signals.

The ECG transient signals are obtained from a “nonlinear system”- a pregnant woman. That’s why we proposed in this paper nonlinear methods of representation (in state space), analysis of recurrences, nonlinear filtering and estimation of parameters (from RP) instead of applying the well known linear methods. With using RP, implementation of real-time processing and separation of maternal and fetal ECG are our future directions.

REFERENCES

- [1] Partha P K., Sarbani P, Goutam S. Fetal ECG Extraction from Single-Channel Maternal ECG Using Singular Value Decomposition. *Med Biol Eng Comput.* 1997 January; 44(1): 51-9.
- [2] Abboud S., Alaluf A., Einav S., Sadeh D. Real-time abdominal fetal ECG recording using a hardware correlator. *Comput Biol Med* 1992; 22 (5): 325-335
- [3] Mooney DM, Groome LJ, Bentz LS, Wilson JD. Computer algorithm for adaptive extraction of fetal cardiac electrical signal, *Proceedings of the 1995 ACM symposium on Applied computing*, p.113-117, February 26-28, 1995, Nashville, Tennessee, United States
- [4] Longini R, Reichert T, Cho J, Crowley J. Near orthogonal basis functions: A real time fetal ECG technique. *IEEE Trans Biomed Eng* 1997; 24:29-43.
- [5] Kam A, Cohen A. Detection of Fetal ECG with IIR Adaptive Filtering and Genetic Algorithms. *IEEE International Conference On Acoustic, Speech, and Signal Processing.* March 15-19, 1999 Civic Plaza, Hyatt Regency - Phoenix, Arizona.
- [6] De Lathauwer L, De Moor B, Vandewalle J. Fetal Electrocardiogram Extraction by Blind Source Subspace Separation. *IEEE Med Biol Eng Comput*, 2000 May; 47(5):567-72.
- [7] Khamene A, Negahdaripour S. A New Method for the Extraction of Fetal ECG from the Composite Abdominal Signal. In: *IEEE Med Biol Eng Comput*, 2000 April; 47(4):507-16.
- [8] Marcus Richter, Thomas Schreiber, Daniel T. Kaplan. Fetal ECG extraction with nonlinear state space projections. *IEEE Trans. Biomed. Eng*; 45: 133-137
- [9] Kantz H, Schreiber T. (2005). *Nonlinear Time Series Analysis*. s.l.: Cambridge University Press.
- [10] Marwan N, *Encounters with Neighbours. Current Developments of Concepts Based on Recurrence Plots and Their Applications*, PhD Thesis, Institut für Physik, Fakultät Mathematik und Naturwissenschaften, Universität Potsdam, 2003.
- [11] M. A. Hasan, M. I. Ibrahimy, M. B. I. Reaz. Fetal ECG Extraction from Maternal Abdominal ECG Using Neural Network. *J. Software Engineering & Applications*, 2009, 2: 330-334
- [12] Birleanu Florin, et al., *Transient Signal Detection Using Recurrence Plot Analysis*. GRETSI 2011, sept., Bordeaux, France.

Data-Intensive Computing with Map-Reduce and Hadoop

Shamil Humbetov
Department of Computer Engineering
Qafqaz University
Baku, Azerbaijan
humbetov@gmail.com

Abstract – Every day, we create 2.5 quintillion bytes of data — so much that 90% of the data in the world today has been created in the last two years alone. This data comes from everywhere: sensors used to gather climate information, posts to social media sites, digital pictures and videos, purchase transaction records, and cell phone GPS signals to name a few. The IDC sizing of the digital universe – information that is either created or captured in digital form and then replicated in 2006 – is 161 Exabyte, growing to 988 Exabyte in 2010, representing a compound annual growth rate (CAGR) of 57%. A variety of system architectures have been implemented for data-intensive computing and large-scale data analysis applications including parallel and distributed relational database management systems which have been available to run on shared nothing clusters of processing nodes for more than two decades. However most data growth is with data in unstructured form and new processing paradigms with more flexible data models were needed. Several solutions have emerged including the MapReduce architecture pioneered by Google and now available in an open-source implementation called Hadoop used by Yahoo, Facebook, and others. 20% of the world’s servers go into huge data centers by the “Big 5” – Google, Microsoft, Yahoo, Amazon, eBay [1].

Index Terms – component, Hadoop, MapReduce, Data Intensive Computing

I. INTRODUCTION

The rapid growth of Internet and WWW has led to vast amounts of information available online. In addition, social, scientific and engineering applications have created large amounts of both structured and unstructured information which needs to be processed, analyzed, and linked [2]–[4]. Nowadays, data-intensive computing typically uses modern data center architectures and massive data processing paradigms.

The requirements for data-intensive analysis of scientific data across distributed clusters or data centers have grown significantly in the recent years. The most widely used frameworks for big data processing are MapReduce, developed by Google, and its open source implementation from Yahoo, Hadoop. Based on Java, it offers a complete ‘infrastructure’ for reliable, scalable and distributed computing. In the MapReduce framework, a computing task can be executing on an arbitrarily large set of nodes, as long

as it has been expressed as a sequence of Maps (independent computations on subsets of the input data) and Reduce (merge of Map results).

The MapReduce framework allows also efficient querying of massive datasets, through specialized, high level querying languages, which are compiled into MapReduce processes.

The MapReduce framework has aligned High Performance Computing (HPC) with the processing of massive datasets. The popularity of its most famous implementation, Hadoop, is driving the emergence of a full ecosystem, which includes query tools, such as Pig and Hive, Zookeeper control libraries, Sqoop data loader, log management (Flume, Scribe) and integration with other NoSQL databases, such as Cassandra.

MapReduce implementations basically work in batch mode. Real-time processing of datasets is, at best, a makeshift job. Initiatives, such as Google’s Percolator or Yahoo’s S4 Distributed Steam Computer Platform, are alternatives towards a real time approach. Their evolution, and possible convergence with the evolution of Hadoop, should be followed closely as they will form the basic platform for advanced realtime data intensive solutions.

Expressing computations as MapReduce processes requires a new way of thinking about computational-intensive algorithms for developers, with a steep learning curve. MapReduce is not suitable for real-time computation and open-source real-time alternatives to Hadoop are still immature [5].

MapReduce is an attractive model for parallel data processing in high performance cluster computing environments. The scalability of MapReduce is proven to be high, because a job in the MapReduce model is partitioned into numerous small tasks running on multiple machines in a large-scale cluster. MapReduce is a widely used method of parallel computation on massive data. MapReduce was designed (by Google, Yahoo, and others) to marshal all the storage and computation resources of a dedicated cluster computer. The most recently published report indicates that, by 2008, Google was running over one hundred thousand MapReduce jobs per day and processing over 20 PB of data in the same period [6]. By 2010, Google had created over ten thousand distinct MapReduce programs performing a variety

of functions, including large-scale graph processing, text processing etc.

Google File System and Hadoop Distributed File System have common design goals. They are both targeted at data intensive computing applications where massive data files are common. Both are optimized in favor of high sustained bandwidths instead of low latency, to better support batch-processing style workloads. Both run on clusters built with commodity hardware components where failures are common, motivating the inclusion of built-in fault tolerance mechanisms through replication.

In both systems, the filesystem is implemented by user level processes running on top of a standard operating system (in the case of GFS, Linux). A single GFS master server running on a dedicated node is used to coordinate storage resources and manage metadata. Multiple slave servers (*chunkserver*s in Google parlance) are used in the cluster to store data in the form of large blocks (*chunks*), each identified with a 64-bit ID. Files are saved by the chunkserver on local disk as native Linux files, and accessed by chunk ID and offset within the chunk. Both HDFS and GFS use the same default chunk size (64MB) to reduce the amount of metadata needed to describe massive files, and to allow clients to interact less often with the single master. Finally, both use a similar replica placement policy that saves copies of data in many locations—locally, to the same rack, and to a remote rack — to provide fault tolerance and improve performance.

II. DISTRIBUTED FILE SYSTEMS

Distributed data intensive computing To store, manage, access, and process vast amount of data represents a fundamental requirement and an immense challenge in order to satisfy needs to search, analyze, mine, and visualize the data and information. Data intensive computing is intended to address this need.

A. Google File System

The Google File System (GFS) is a proprietary Distributed File System developed by Google. It is designed (Figure 1) to provide efficient, reliable access to data using large clusters of commodity hardware.

The files are huge and divided into chunks of 64 megabytes [7]. Most files are mutated by appending new data rather than overwriting existing data: once written, the files are only read and often only sequentially. This DFS is best suited for scenarios in which many large files are created once but read many times. The GFS is optimized to run on computing clusters where the nodes are cheap computers. Hence, there is a need for precautions against the high failure rate of individual nodes and data loss. In the Google file system there can be 100 to 1000 PCs in a cluster can be used.

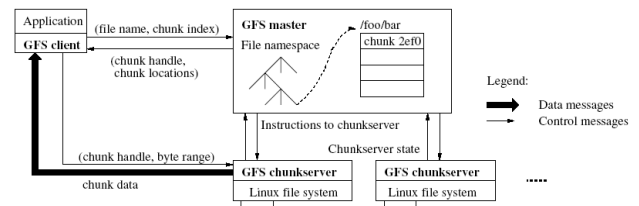


Fig. 1. Google File System

1) Chunkserver Architecture

Server

- Stores 64 MB file chunks on local disk using standard Linux filesystem, each with version number and checksum
- Read/write requests specify chunk handle and byte range
- Chunks replicated on configurable number of chunkserver (default: 3)
- No caching of file data

2) Client

- Issues control (metadata) requests to master server
- Issues data requests directly to chunkserver
- Caches metadata
- Does no caching of data
 - No consistency hence difficulties among clients
 - Streaming reads (read once) and append writes (write once) don't benefit much from caching at client

B. Hadoop Distributed File System

HDFS, the Hadoop Distributed File System, is a distributed file system designed (Figure 2) to hold very large amounts of data (terabytes or even petabytes), and provide high throughput access to this information. Files are stored in a redundant fashion across multiple machines to ensure their durability to failure and high availability to very parallel applications.

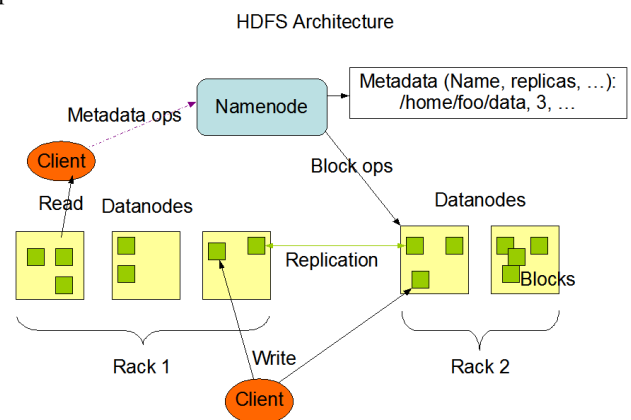


Fig. 2. Hadoop Distributed File System

HDFS has a master /slave architecture. An HDFS cluster consists of a single NameNode, a master server that manages the file system namespace and regulates access to files by clients. In addition, there are a number of DataNodes, usually

one per node in the cluster, which manages storage attached to the nodes that they run on. Internally, a file is split into one or more blocks and these blocks are stored in a set of DataNodes. The NameNode executes file system namespace operations like opening, closing, and renaming files and directories. It also determines the mapping of blocks to DataNodes. The DataNodes are responsible for serving read and write requests from the file system's clients. The DataNodes also perform block creation, deletion, and replication upon instruction from the NameNode [8].

III. MAPREDUCE AS A DATA PROCESSING MODEL

MapReduce is also a data processing model. Its greatest advantage is the easy scaling of data processing over multiple computing nodes. Under the MapReduce model, the data processing primitives are called mappers and reducers. In the mapping phase, MapReduce takes the input data and feeds each data element to the mapper. In the reducing phase, the reducer processes all the outputs from the mapper and arrives at a final result. In simple terms, the mapper is meant to filter and transform the input into something that the reducer can aggregate over [9].

Before developing the MapReduce framework, Google used hundreds of separate implementations to process and compute large datasets. Most of the computations were relatively simple, but the input data was often very large. Hence the computations needed to be distributed across hundreds of computers in order to finish calculations in a reasonable time. MapReduce is highly efficient and scalable, and thus can be used to process huge datasets. When the MapReduce framework was introduced, Google completely rewrote its web search indexing system to use the new programming model. The indexing system produces the data structures used by Google web search. There is more than 20 Terabytes of input data for this operation. At first the indexing system ran as a sequence of eight MapReduce operations, but several new phases have been added since then. Overall, an average of hundred thousand

MapReduce jobs is run daily on Google's clusters, processing more than twenty Petabytes of data every day. The idea of MapReduce is to hide the complex details of parallelization, fault tolerance, data distribution and load balancing in a simple library. In addition to the computational problem, the programmer only needs to define parameters for controlling data distribution and parallelism. Like Google's MapReduce, Hadoop uses many machines in a cluster to distribute data processing. The parallelization doesn't necessarily have to be performed over many machines in a network. There are different implementations of MapReduce for parallelizing computing in different environments. Hadoop is a distributed file system that can run on clusters ranging from a single computer up to many thousands of computers. Hadoop was inspired by two systems from Google, MapReduce and Google File System.

IV. HADOOP – PROCESSING LARGE AMOUNTS OF DATA

Hadoop – is the System for processing mind-boggling large amounts of data [10]. Hadoop got its start in Nutch. We live in the data age. It's not easy to measure the total volume of data stored electronically, but an IDC estimate put the size of the "digital universe" at 0.18 zettabytes in 2006, and is forecasting a tenfold growth by 2011 to 1.8 zettabytes.* A zettabyte is 10^{21} bytes, or equivalently one thousand exabytes, one million petabytes, or one billion terabytes. That's roughly the same order of magnitude as one disk drive for every person in the world. Hadoop can run MapReduce programs written in various languages; Java, Ruby, Python, and C++. Most important, MapReduce programs are inherently parallel, thus putting very large-scale data analysis into the hands of anyone with enough machines at their disposal. There are Hadoop clusters running today that store petabytes of data. Scaling Hadoop to 4000 nodes at Yahoo!

Hadoop is good at processing large amount of data in parallel. The idea is to breakdown the large input into smaller chunks and each can be processed separately on different machines. That way, we can alleviate the IO bottleneck across many machines to achieve better overall performance. The infrastructure has abstracted you out from the complexity of distributed computing. So, the user has no worry of machine failure, data availability and coordination.

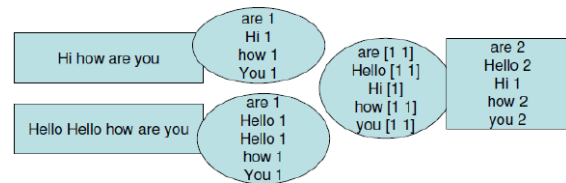


Fig. 3. MapReduce

GFS is a scalable distributed file system for data-intensive applications. GFS provides a fault tolerant way to store data on commodity hardware and deliver high aggregate performance to a large number of clients.

MapReduce is a toolkit (Figure 3) for parallel computing and is executed on a large cluster of commodity machines. There is quite a lot of complexity involved when dealing with parallelizing the computation, distributing data and handling failures. Using MapReduce allows users to create programs that run on multiple machines but hides the messy details of parallelization, fault-tolerance, data distribution, and load balancing. MapReduce conserves network bandwidth by taking advantage of how the input data is managed by GFS and is stored on the local storage device of the computers in the cluster.

MapReduce and the Hadoop File System (HDFS), which is based on GFS, what defines the core system of Hadoop. MapReduce provides the computation model while HDFS provides the distributed storage. MapReduce and HDFS are designed to work together. While MapReduce is taking care of the computation, HDFS is providing high throughput of

data. Hadoop has one machine acting as a NameNode server, which manages the file system namespace. All data in HDFS are split up into block sized chunks and distributed over the Hadoop cluster. The NameNode manages the block replication. If a node has not answered for some time, the NameNode replicates the blocks that were on that node to other nodes to keep up the replication level.

Most nodes in a Hadoop cluster are called DataNodes; the NameNode is typically not used as a DataNode, except for some small clusters. The DataNodes serve read/write requests from clients and perform replication tasks upon instruction by NameNode. DataNodes also run a TaskTracker to get map or reduce jobs from JobTracker. The JobTracker runs on the same machine as NameNode and is responsible for accepting jobs submitted by users. The JobTracker also assigns Map and Reduce tasks to Trackers, monitors the tasks and restarts the tasks on other nodes if they fail.

V. MAPREDUCE

MapReduce programming consists of writing two functions, a map function, and a reduce function. The map function takes a key, value pair and outputs a list of intermediate values with the key. The map function is written in such a way that multiple map functions can be executed at once, so it's the part of the program that divides up tasks. The reduce function then takes the output of the map functions, and does some process on them, usually combining values, to generate the desired result in an output file.

Figure 4 below shows a picture representing the execution of a MapReduce job [11].

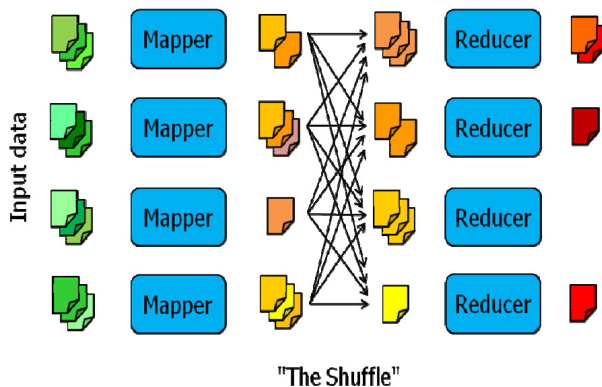


Fig. 4. MapReduce Job

When a MapReduce program is run by Hadoop, the job is sent to a master node, the jobtracker, which has multiple "slave" nodes, or tasktrackers that report to it and ask for new work whenever they are idle. Using this process, the jobtracker divides the map tasks (and quite often the reduce tasks as well) amongst the tasktrackers, so that they all work in parallel. Also, the jobtracker keeps track of which

tasktrackers fail, so their tasks are redistributed to other task trackers, only causing a slight increase in execution time. Furthermore, in case of slower workers slowing down the whole cluster, any tasks still running once there are no more new tasks left are given to machines that have finished their tasks already. Not every process nodes have a small piece of a larger file, so that when a file is accessed, the bandwidth of a large number of hard disks is able to be utilized in parallel. In this way, the performance of Hadoop may be able to be improved by having the I/O of nodes work more concurrently, providing more throughput.

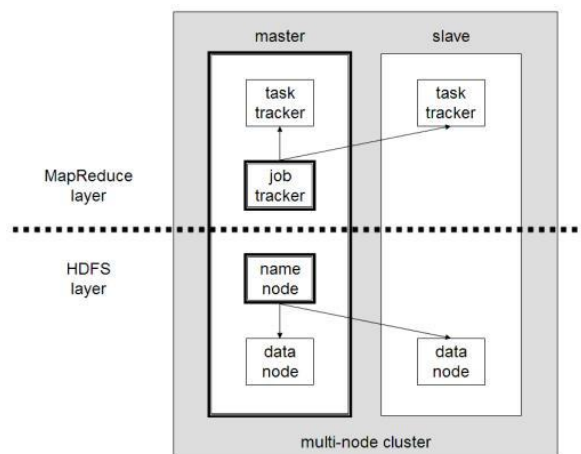


Fig. 5. Hadoop Cluster Architecture

Map Reduce works in the following manner in below 7 tasks:

1. The Map-Reduce library in the user program first splits the input into M pieces of typically 16 megabytes to 64 megabytes (MB) per piece. It then starts up many copies of the program on a cluster of machines. (Refer Figure 5)
2. One of the copies of the program is special- the master copy. The rest are workers that are assigned work by the master. There are M map task and R reduce tasks to assign; the master picks idle workers and assign each one a task
3. A worker who is assigned a map task reads the contents of the contents of the corresponding input split. It parses key/value pairs out of the input data and passes each pair to the user-defined Map function. The intermediate key/value pairs produced by the Map function are buffered in memory.
4. Periodically, the buffered pairs are written to local disk partitioned into R regions by the partitioning function. The location of these buffered pairs on the local disk are passed back to the master, who is responsible for forwarding these locations to the reduce workers
5. When a reduce worker is modified by the master about these locations. it uses remote procedure calls to read buffered data from the local disk of map workers. When a reduce worker has read all intermediate data, it sorts it by the

intermediate keys. The sorting is needed because typically many different key map to the same reduce task.

6. The reduce worker iterate over the sorted intermediate data and for each unique key encountered, it passes the key and the corresponding set of intermediate values to the user's Reduce function. The output of the Reduce function is appended to the final output file for this reduce partition

7. When all map task and reduce task have been completed, the master wakes up the user program. At this point, the Map-Reduce call in the user program returns back to the user code.

A. Theoretical Representation of Map Reduce

1. Data are represented as a <key, value> pair
2. Map: <key, value> → multiset of <key, value> pairs user defined, easy to parallelize
3. Shuffle: Aggregate all <key, value> pairs with the same key. executed by underlying system
4. Reduce: <key, multiset(value)> → <key, multiset(value)> user defined, easy to parallelize
5. Can be repeated for multiple rounds [12].

Map Reduce works as a Job Tracker and Task Tracker.

• Map/Reduce Master “Jobtracker”

- _ **Accepts** Map-Reduce jobs submitted by users
- _ **Assigns** Map and Reduce tasks to Tasktrackers
- _ **Monitors** task and tasktracker status, re-executes tasks upon failure

• Map/Reduce Slaves “Tasktrackers”

- _ **Run** Map and Reduce tasks upon instruction from the Jobtracker
 - _ **Manage** storage and transmission of intermediate output.
- Job tracker functions in the following manner:-

- Handles all jobs
- Makes all scheduling decisions
- Breaks jobs into tasks, queues up
- Schedules tasks on nodes close to data
- Location information comes from InputSplit
- Monitors tasks
- Kills and restarts tasks if they fail/hang/disappear
- Task tracker works in the following manner:-

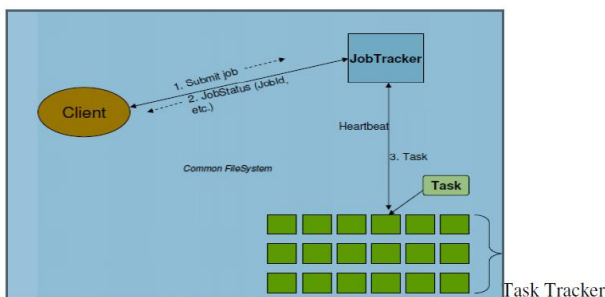


Fig. 6. Parallel MapReduce Computations

- Asks for new tasks, executes, monitors and reports status

The programmer can be mostly oblivious to parallelism and distribution; the programming model readily enables parallelism, and the MapReduce implementation takes care of the complex details of distribution such as load balancing, network performance and fault tolerance. The programmer has to provide parameters for controlling distribution and parallelism, such as the number of reduce tasks to be used which is described in the later part of this paper by referring the example. (Figure 6) Defaults for the control parameters may be inferable. In this section, I have made the clarification on the opportunities for parallelism in a distributed execution of MapReduce computations.

VI. REFERENCES

- [1] Big Data Platform [Online]. Available: <http://www-01.ibm.com/software/data/bigdata/>
- [2] F. Berman, "Got data?: a guide to data preservation in the information age," *Commun. ACM*, vol. 51, pp. 50–56, December 2008. [Online]. Available: <http://doi.acm.org/10.1145/1409360.1409376>
- [3] "Data intensive computing." [Online]. Available: http://en.wikipedia.org/wiki/Data_Intensive_Computing
- [4] L. Wang, M. Kunze, J. Tao, and G. von Laszewski, "Towards building a cloud for scientific applications," *Advances in Engineering Software*, vol. 42, no. 9, pp. 714–722, 2011.
- [5] http://ascentlookout.atos.net/en-us/enabling_information_technologies/big_data_processing/default.htm
- [6] Jeffrey Dean and Sanjay Ghemawat. Mapreduce: simplified data processing on large clusters. *Commun. ACM*, 51(1):107–113, 2008.
- [7] Review of Distributed File Systems: Concepts and Case Studies ECE 677 Distributed Computing Systems
- [8] HDFS Architecture Guide [Online] Available: http://hadoop.apache.org/docs/current/hdfs_design
- [9] Hadoop in Action – Chuck Lam
- [10] APACHE HADOOP - PETABYTES AND TERAWATTS [Online]. Available: <http://www.youtube.com/watch?v=SS27F-hYWfU&feature=related>
- [11] Levy E. and Silberschatz A., "Distributed FileSystems: Concepts and Examples"
- [12] Yahoo Research- <http://dimacs.rutgers.edu/Workshops/Parallel/slides/suri.pdf>

SESSION 2

Communication, network and hardware

AICT 2012

Web based system for the bee colony remote monitoring

Aleksejs Zacepins, Toms Karasha
Latvia University of Agriculture
Faculty of Information Technologies
Liela iela 2, Jelgava, Latvia
E-mail: alzpostbox@gmail.com; toms.karasha@gmail.com;

Abstract - Monitoring of bioprocesses is an important problem and challenge for researchers and information technology specialists. Beekeeping is one of the agricultural sub directions where monitoring technics and methods can be adapted and applied. Integration of information technologies into beekeeping process can improve the beekeepers knowledge about behavior of individual bee colonies. One of the objectives in the field of bee monitoring is to develop real time on-line tools for continuous observations of bees during their life and production using the automatic solutions avoiding exposure of bees to additional stress or unproductive activities. The aim of these technical tools is not to replace but rather to support the beekeeper. The aim of this paper is to describe developed web based system for real time bee monitoring.

I. INTRODUCTION

In Europe the apiculture is an important factor for human society, economy and ecology. Honey is produced on industrial, semi industrial and on a private level and used as food and as additives for medical and pharmaceutical products. There are two sides of the economics of apiculture. On the one side it is a considerable market for honey products and on the other side is the important role of honeybees in pollinating agricultural crops and natural plants. It has been estimated that about 35% of the food production depend directly or indirectly on insect animal pollination and that 84% of crop species depend at least to some extent upon it [1]. Although the direct economic value in the EU is only around 228 million €, the indirect value was estimated for the EU (2005) to be 14.2 billion €. Taking into consideration that the numbers of wild pollinators are declining the importance of apiculture is greater today than ever.

Despite its significance for the European agriculture the beekeeping sector is in decline. There are several reasons for this like land-use intensification, pesticide poisoning, colony diseases and parasites. All this reasons can cause colony deaths, which could be enough for small-scale hobbyists to stop beekeeping, but are also a crucial limitation for the economic survival of (semi) professional producers. Apiculture is a profession with high investments in material and

equipment, which depends on a forecastable economic income. Therefore unforeseeable colony losses put the enterprises on the brink of bankruptcy. Especially in the Eastern European countries which have no state support system for this kind of losses many beekeepers had to abandon their apiaries.

Despite large scientific efforts taken by the European nations to determine the causes for the rising numbers of colony losses, there are still no practical solutions for this apiculture to prevent them. What the beekeepers need is a cheap and reliable automatic diagnostic tool, which alerts them about possible threats to their hives, so that they can take action to prevent the colony losses. This can be achieved by developing an automatic system, which can determine a healthy state of the colony and can alert the farmer if any aberrations from this state are metered.

The main aim of this paper is to describe developed web based system for monitoring the honeybee colony activity based on temperature measurements

II. DEVELOPED TEMPERATURE MEASUREMENT SYSTEM

Applying information technologies it is possible to develop the real time measurement system to continuously monitor bee colonies during their life and production stage using the automatic, automated and communication technology based solutions, without exposing the bees to avoidable stress and waste of resources, thus detecting different states and health status of colonies and apiaries more generally enabling rapid reaction by the beekeeper in the case of necessity.

Generally several parameters of individual colonies can be continuously automatically measured and/or analyzed: temperature by temperature sensors [2–8], temperature by infrared imaging [6], [9–11], air humidity [7], [8], [12], [13] gas content [3], [14] sound [15–17], vibration of the beehive [18], counting of outgoing and incoming bees [19], video observation [8], [20], weighing of the colony [21], [22] and others. But costs of those measurements are very different thus limiting economically feasible solutions.

Nowadays rapid improvements in temperature sensing overall allow economically feasible applications in beekeeping. Continuous data capture and analysis can be used to monitor individual bee hives and the data can be adapted for individual bee colony maintenance. Bee hive temperature measurements can be used to understand and to monitor the honey bee colony activity and changes in its behavior [23].

Nowadays temperature sensors are cheap, small and it is easy to install them. But temperature sensor usage in beekeeping is not the newest scientific direction.

Activity and behavior features of bees have been as investigation object for many researches, starting from far 1926 when W.E.Dunham measured temperature inside one bee hive using 8 thermometers [24].

Our research was held in Jelgava, Latvia when using one wire based temperature measurement system bee hives were remotely monitored. Each hive was equipped with small digital temperature sensor. All sensors were sequentially connected with Temp08 interface device, while it was connected with the end PC using the COM port [25] (see Fig.1 and Fig.2).



Fig.2. Developed bee hive temperature measurement system in real circumstances

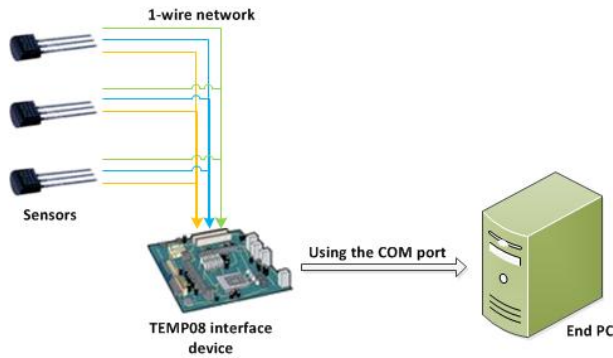


Fig.1. Hardware architecture of the temperature measurement system

III. DEVELOPED APPLICATIONS AND WEB INTERFACE

Technically developed temperature measurement system is only a part of the complex system. Another not the less important part is a software.

All of the bee colony temperature measurements should be not only demonstrated in real time to the beekeeper but also saved for their future analysis. Therefore it is necessary to develop PC application for temperature data storage and demonstration.

The whole software system architecture is demonstrated in Fig.3:

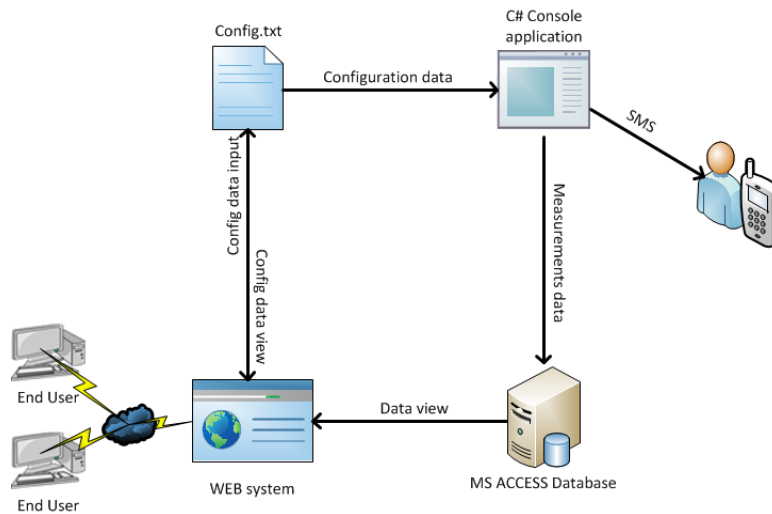


Fig.3. Developed software system architecture for bee colony temperature monitoring

The software system consists of 5 separate parts:

1. **Configuration file** – it is a .txt file which is needed for basic parameters initialization, which is used by the console application.

Defined parameters are:

- Number of the COM port – initialization of the PC COM port which is used for connecting the Temp08 device for data transmitting procedure;
- Time thread – time in seconds, which is needed to complete the data transmitting procedure;
- Count of bits – complete number of bits, which are transmitted;
- Serial number of sensor for error reporting – initialization of one exact sensor for error reporting operation
- Maximal value of the temperature – definition of maximal allowed limit of one specific sensor. When real value reaches the maximal limit, error notification is sent.
- Minimal value of the temperature – definition of minimal allowed limit of one specific sensor. When real value reaches the minimal limit, error notification is sent.
- E-mail address – e-mail, where to send error notification message.
- Backup – a path where to copy measurement database.

2. **Database** – is developed MS Access database where to save all temperature data from sensors. Database consists only from two tables: sensor table, where all sensors are defined and measurement table. Measurement table consist of three fields: sensor number, time moment when measurement is taken and temperature data to that moment.

Table I
Measurement table

Sensor serial number	Date and time	Result
3C000X4FGJ	10.05.2012 14:45:34	8,56

3. **Console application** – is used to save all the measurements from the sensors to MS Access database, as well for error reporting procedure.

Authors have made experiments to find out how fast it is possible to transmit the data from sensors to the end PC.

Table II
One sensor reading data in seconds

Sensors 1	1	2	3	4	5
	1,81	2,01	2,11	1,56	1,87
	6	7	8	9	10
	2,17	1,99	1,55	1,82	1,56

Table is demonstrating time in seconds for all sensor data transmitting that were used during authors practical experiment.

Table III
23 sensors reading data in seconds

Sensors 23	1	2	3	4	5
	26	25,8	26,3	26,3	25,5
	6	7	8	9	10
	26	25,8	26,5	26,7	25,2

4. **Web interface** – the main part of the software system. Is used for easy access to all measurements. The main idea is that a beekeeper from any place, where there is Internet coverage could connect to the web server and check temperature in the hives. As beekeepers usually are not very experienced in using PC applications, WEB interface should be as simple as possible, without any possibility to make any wrong actions. Web system is developed using the ASP.Net technology. Developed web interface is demonstrated in Fig.4, Fig.5:

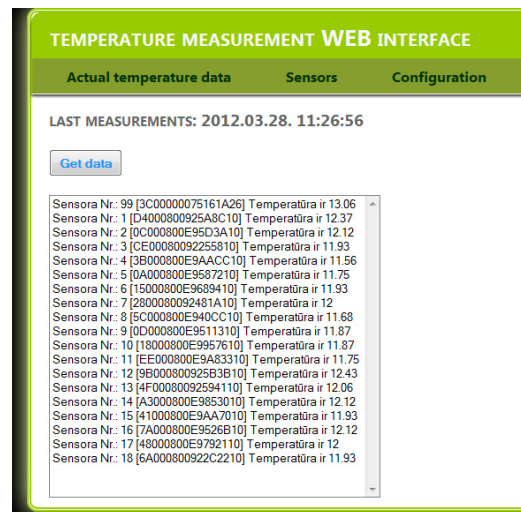


Fig.4. Developed WEB interface

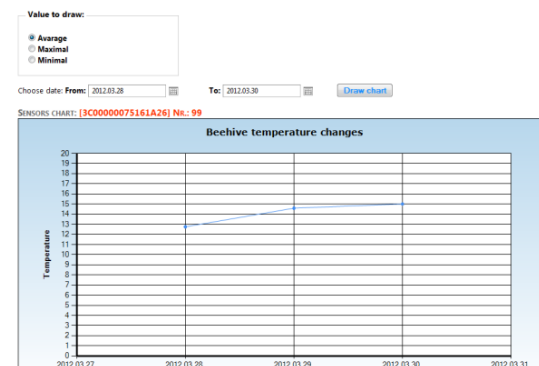


Fig.5. Analysis module of the WEB interface

Web interface is divided into 3 modules (forms):
 1 module – actual information about bee hives temperature (last measurements for all sensors).
 2 module – option for detailed analysis of bee hives temperature with option to graphically demonstrate average, minimal or maximal temperatures for hive.
 3 module – option for administration and modifying information in configuration file.

5. **Web server configuration** – to publish web system it is needed to use some web server, in authors case integrated windows IIS (Internet Information Service) were used.

IV. CONCLUSIONS

The developed software system is used to store all the measurement data obtained from temperature sensors. All temperature data is stored on the PC for future data analysis and decision making. The temperature measurements are used to determine bee colony activity and behavior features.

WEB options are needed to give the beekeeper access to data in real time. System in addition has error reporting option for informing the beekeepers about unusual situations in bee colonies.

The developed temperature measurement system in combination with developed software could be used also in other agricultural fields where it is needed to make some precise decision based on temperature data, for example temperature measurements in greenhouses.

ACKNOWLEDGMENT

Academic study and publication financed by the project „Support for doctoral studies in LUA” / 2009/0180/1DP/1.1.2.1.2/09/IPIA/VIAA/017/ agreement Nr. 04.4-08/EF2. D1.09.

REFERENCES

[1] K. Delaplane and D. F. Mayer, *Crop pollination by bees*. Cabi, 2000, p. 345.
 [2] L. Fahrenholz, I. Lamprecht, and B. Schrickler, “Thermal investigations of a honey bee colony: thermoregulation of the hive during summer and winter and heat production of members of different bee castes,” *Journal of comparative physiology. B, Biochemical, systemic, and environmental physiology*, vol. 159, no. 5, pp. 551-560, 1989.
 [3] K. Van Nerum and H. Buelens, “Hypoxia-Controlled Winter Metabolism in Honeybees (*Apis mellifera*),” *Comparative Biochemistry and Physiology Part A: Physiology*, vol. 117, no. 4, pp. 445-455, Aug. 1997.
 [4] B. Chuda-Mickiewicz and J. Prabucki, “Temperature in winter cluster bee colony wintering in a hive of cold comb arrangement,” *Pszczelnicze Zestyty Naukowe*, vol. 40, no. 2, pp. 71-79, 1996.
 [5] E. Stalidzans, V. Bilinskis, and A. Berzonis, “Determination of development periods of honeybee colony by temperature in hive in latvia, year 2000,” *Apiacta*, pp. 4-8, 2002.

[6] E. K. Eskov and V. A. Toboev, “Seasonal dynamics of thermal processes in aggregations of wintering honey bees (*Apis mellifera*, Hymenoptera, Apidae),” *Entomological Review*, vol. 91, no. 3, pp. 354-359, Jun. 2011.
 [7] O. C. Vornicu and I. Olah, “Monitoring System of Bee Families Activity,” in *7th International Conference on Development and Application Systems*, 2004, pp. 88-94.
 [8] J. Meitalovs, A. Histjajevs, and E. Stalidzans, “Automatic Microclimate Controlled Beehive Observation System,” in *8th International Scientific Conference “Engineering for Rural Development,”* 2009, pp. 265-271.
 [9] E. K. Eskov and V. A. Toboev, “Mathematical modeling of the temperature field distribution in insect winter clusters,” *Biophysics*, vol. 54, no. 1, pp. 85-89, 2009.
 [10] M. Kleinhenz, B. Bujok, S. Fuchs, and J. Tautz, “Hot bees in empty broodnest cells: heating from within,” *Journal of Experimental Biology*, vol. 206, no. 23, pp. 4217-4231, Dec. 2003.
 [11] J. A. Shaw, P. W. Nugent, J. Johnson, J. J. Bromenshenk, C. B. Henderson, and S. Debnam, “Long-wave infrared imaging for non-invasive beehive population assessment,” *Optics Express*, vol. 19, no. 1, pp. 399-408, 2011.
 [12] H. Human, S. W. Nicolson, and V. Dietemann, “Do honeybees, *Apis mellifera* scutellata, regulate humidity in their nest?,” *Die Naturwissenschaften*, vol. 93, no. 8, pp. 397-401, Aug. 2006.
 [13] B. Kraus and H. H. W. Velthuis, “High Humidity in the Honey Bee (*Apis mellifera* L.) Brood Nest Limits Reproduction of the Parasitic Mite *Varroa jacobsoni* Oud.,” *Naturwissenschaften*, vol. 84, no. 5, pp. 217-218, May 1997.
 [14] E. E. Southwick, D. W. Roubik, and J. M. Williams, “Comparative energy balance in groups of africanized and European honey bees: ecological implications,” *Comparative Biochemistry and Physiology*, vol. 97A, no. 1, pp. 1-7, 1990.
 [15] E. K. Eskov and V. A. Toboev, “Analysis of statistically homogeneous fragments of acoustic noises generated by insect colonies,” *Biophysics*, vol. 55, no. 1, pp. 92-103, Jun. 2010.
 [16] S. Ferrari, M. Silva, M. Guarino, and D. Berckmans, “Monitoring of swarming sounds in bee hives for early detection of the swarming period,” *Computers and Electronics in Agriculture*, vol. 64, no. 1, pp. 72-77, Nov. 2008.
 [17] D. G. Dietlein, “A method for remote monitoring of activity of honeybee colonies by sound analysis,” *Journal of Apicultural Research*, vol. 24, no. 3, pp. 176-183, 1985.
 [18] M. Bencsik, J. Bencsik, M. Baxter, A. Lucian, J. Romieu, and M. Millet, “Identification of the honey bee swarming process by analysing the time course of hive vibrations,” *Computers and Electronics in Agriculture*, vol. 76, no. 1, pp. 44-50, Mar. 2011.
 [19] C. Liu, J. J. Leonard, and J. J. Feddes, “Automated monitoring of flight activity at a beehive entrance using infrared light sensors,” *Journal of Apicultural Research*, vol. 29, no. 1, pp. 20-27, 1990.
 [20] J. Campbell, L. Mummert, and R. Sukthakar, “Video Monitoring of Honey Bee Colonies at the Hive Entrance,” *Visual observation & analysis of animal & insect behavior, ICPR*, vol. 8, pp. 1-4, 2008.
 [21] T. D. Seeley and P. K. Visscher, “Survival of honeybees in cold climates: the critical timing of colony growth and reproduction,” *Ecological Entomology*, vol. 120, no. 1, pp. 826-88, Feb. 1985.
 [22] J. Nickelson, “HoneyBeeNet,” *NASA*, 2010. .
 [23] A. Zacepins, “Application of bee hive temperature measurements for recognition of bee colony state,” in *Applied Information and Communication Technologies*, 2012, pp. 216-221.
 [24] W. Dunham, “Hive temperatures for each hour of a day,” *Ohio J. Sci*, pp. 181-188, 1926.
 [25] A. Zacepins, J. Meitalovs, V. Komasilovs, and E. Stalidzans, “Temperature sensor network for prediction of possible start of brood rearing by indoor wintered honey bees,” in *2011 12th International Carpathian Control Conference (ICCC)*, 2011, pp. 465-468.

High Speed Digital Filter Design using Register Minimization Retiming and Parallel Prefix Adders

Deepa Yagain¹, Dr. Vijaya Krishna A.²
Department of E&C (VLSI Design & Embedded Systems)
People's Education Society Institute of Technology
Bangalore-560 085, Karnataka, INDIA
¹deepa.yagain@gmail.com, ²vijaykrishna@gmail.com

Abstract—Design of complex filter solutions from the solution space can be handled at different levels in VLSI design process. Thus, finding reasonably good solutions can consume lot of design time and effort. Hence optimization and automation of filter designs can reduce the design time as well as increase the design performance. In this paper, a Design optimization platform is designed such that synthesizable RTL is obtained from input such as Data Flow Graphs (DFGs) for any digital filters. The optimization is performed using Register minimization Retiming. While synthesizing, architectural optimisations like usage parallel prefix adders is done. The input specifications in the current work are taken in the form of Matrices which are derived from DFGs. Retiming using register minimization is a process in which the location of the registers is altered in such a way that the overall clock period reduces, thereby increasing the clock frequency. This happens due to reduction in the critical path which bounds the speed of the design. Due to intelligent placement of registers in the register minimization retiming, the number of registers gets minimized there by minimizing the area .

Since all the Digital filters are made of adders, multipliers and delay elements, optimizing these will in turn increase the design performance. Instead of ripple adders, designs make use of parallel prefix adders. It is found that the combinational path delay of the parallel prefix Kogge-Stone Ling adder is much less which can further reduce the clock period and increase the design speed.

Keywords—Register minimisation Retiming, Parallel Prefix Adders, High-Level Synthesis (HLS), Inequality, Floyd-Warshall algorithm, Data Flow Graphs, Linear Programming

INTRODUCTION

Due to VLSI design Complexity in multimillion designs, advanced Design optimization platforms are needed to handle. Also VLSI designers are asked to design circuits with higher performance and no defects. This needs to be done with lesser design time. CAD tools play a pivotal role in achieving these requirements [1][2]. Further to obtain high performance designs in terms of Speed, Area and Power can further increase the need of CAD tools. In ASIC design process [3] starts with set system requirements. From these specifications, high level functional blocks are obtained. These can be later used for obtaining circuit level devices. Present work considers the generation of synthesizable, optimised digital

filters automatically by using Register minimisation retiming technique. The High Level Transformations operate on the functional blocks generated by specifications and alter their performance parameters like Speed, Area and power [4][5][6][7]. The high level transformation used in this paper is Register minimization Retiming. This is chosen as the high level transformation for digital design for the increased quality of design speed and reduction in the area when compared to other transformation methods. This retiming also preserves the design functionality. In literature, many techniques have been developed to attempt an increase in the performance of the circuits at the transistor level [8]. However, due to increased sampling rate constraints of the real time DSP applications like speech compression, telecommunication and data processing, we need much more efficient techniques for enhancing the circuit performance. Every node in the design has several output edges carrying the same signal value. The number of registers required to implement these edges is the maximum number of registers on the particular edge. Register minimization retiming not only attempts to increase the clock rate by critical path reduction but also finds a solution which uses minimum number of registers. This technique along with other optimization techniques like the usage of parallel prefix adders and Voltage scaling are automated to generate a synthesizable HDL which reduces most of the design time and designer can effectively search the solution space to obtain the global optimum solution. This work proposes such a Design optimization platform embedded in MATLAB/Simulink which is a very comprehensive and easy-to-use graphical platform.

I. DESIGN AND ANALYSIS

The input to the designed Design optimization platform is given as Data Flow Graphs. DFGs represent iterations in any DSP Filter blocks. DFGs contain precedence information indicating operation execution sequence. This can also indicate the critical operations which are nothing but operations along the critical path of DFG. This ability of a DFG to encapsulate any Digital Filter without describing the hardware implementation details has made it a better form of algorithm representation [9].

A DFG is a directed graph $G(V, E)$ with set of nodes or vertices V and set of edges E . The set of nodes V are

subdivided into computational nodes, input and output nodes, and conditional nodes. Figure 1 shows the circuit and the Data flow graph of the 3 Tap IIR filter.

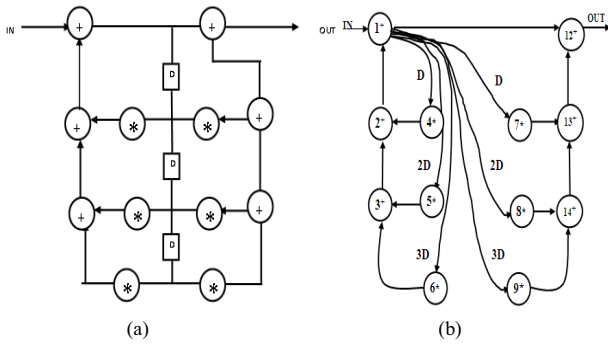


Figure 1. (a) 3 Tap IIR Filter (b) DFG of 3 Tap IIR Filter

Here DFGs can also be represented in the form of matrices. In the present work, two matrices are used for every DFG.

1) Node-weight Matrix -a matrix of all nodes in the graph and the corresponding node weights.

2) Incidence Matrix -defining the edge weights between all the nodes.

Here one matrix gives the connectivity information and another matrix gives the time required by each node to perform certain operation. For any digital filters, majority of the operations include addition and/or multiplication. The input to the Design optimization platform is given in the form of matrices which are extracted from DFGs. Depending on connectivity information and node set of DFGs we obtain the below matrices for 3 Tap IIR filter as shown in Figure 2. Here multiplication operation is assumed to take 2 pipeline stages and addition is assumed to take one pipeline stage.

$$a = [1 \ 1; 2 \ 1; 3 \ 1; 4 \ 2; 5 \ 2; 6 \ 2; 7 \ 2; 8 \ 2; 9 \ 2; 10 \ 1; 11 \ 1; 12 \ 1]$$

$$f = \begin{bmatrix} \text{inf} & \text{inf} & \text{inf} & 1 & 2 & 3 & 1 & 2 & 3 & \text{inf} & \text{inf} & 0; \\ 0 & \text{inf} & \text{inf} & \text{inf} & \text{inf} & \text{inf} & \text{inf} & \text{inf} & \text{inf} & \text{inf} & \text{inf} & \text{inf}; \\ \text{inf} & 0 & \text{inf} & \text{inf} & \text{inf} & \text{inf} & \text{inf} & \text{inf} & \text{inf} & \text{inf} & \text{inf} & \text{inf}; \\ \text{inf} & \text{inf} & 0 & \text{inf} & \text{inf} & \text{inf} & \text{inf} & \text{inf} & \text{inf} & \text{inf} & \text{inf} & \text{inf}; \\ \text{inf} & \text{inf} & 0 & \text{inf} & \text{inf} & \text{inf} & \text{inf} & \text{inf} & \text{inf} & \text{inf} & \text{inf} & \text{inf}; \\ \text{inf} & \text{inf} & \text{inf} & \text{inf} & \text{inf} & \text{inf} & \text{inf} & \text{inf} & \text{inf} & \text{inf} & 0 & \text{inf}; \\ \text{inf} & \text{inf} & \text{inf} & \text{inf} & \text{inf} & \text{inf} & \text{inf} & \text{inf} & \text{inf} & \text{inf} & 0 & \text{inf}; \\ \text{inf} & \text{inf} & \text{inf} & \text{inf} & \text{inf} & \text{inf} & \text{inf} & \text{inf} & \text{inf} & \text{inf} & 0 & \text{inf}; \\ \text{inf} & \text{inf} & \text{inf} & \text{inf} & \text{inf} & \text{inf} & \text{inf} & \text{inf} & \text{inf} & \text{inf} & 0 & \text{inf}; \\ \text{inf} & \text{inf} & \text{inf} & \text{inf} & \text{inf} & \text{inf} & \text{inf} & \text{inf} & \text{inf} & \text{inf} & \text{inf} & 0; \\ \text{inf} & \text{inf} & \text{inf} & \text{inf} & \text{inf} & \text{inf} & \text{inf} & \text{inf} & \text{inf} & \text{inf} & \text{inf} & \text{inf} \end{bmatrix}$$

Figure 2. Matrices of 3 tap IIR Filter

Critical path delay is defined as the sum of delays of all the nodes (combinational circuits) in the critical path. Critical paths are applicable for Directed Acyclic Graphs (DAG). The chain of dependencies that takes the longest time for

computation is called critical path. This determines the operating clock frequency.

The approach used for the calculation of critical path is:

- Input the node-weight matrix and the incidence matrix.
- Find all the 0-weight edges and form a matrix of their source and destination nodes. 0-weight indicated that there is no delay element in between two nodes.
- For each row in the matrix, if destination node of any 0-weight edge path is same as the source node of the 0-weight edge path, then the path considered is zero. Search for nodes with zero weights and Repeat this to obtain a matrix whose rows will have the nodes of all possible 0-weight edge paths in the graph.
- Calculate computational time of each 0-weight edge path from this matrix.
- Find the 0-weight edge path with the greatest computational time. This is the critical path and its computational time is the critical path delay

The generated graph and critical path for the considered 3 Tap IIR filter is given in Figure 3.

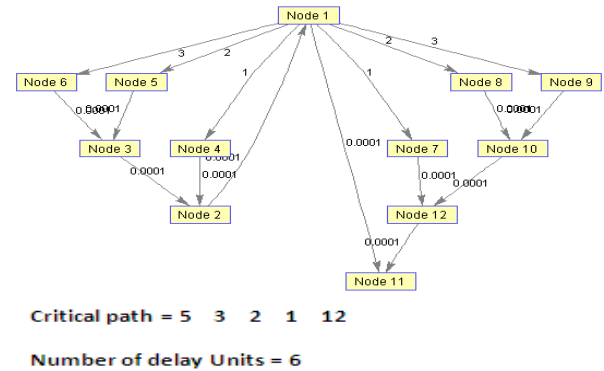


Figure 3. DFG and critical path generated by designed retiming Design optimization platform 3 Tap IIR Filter

Given a DFG, $G = (V, E)$ with a weight function $w: E \rightarrow R$, where R is the set of real numbers, we need determine the weight of the shortest path between all pairs of vertices in G . In Retiming this is needed for knowing whether constraint graph has negative cycles for solving the inequalities of the graph. Shortest path algorithms can be used to perform this task. In this particular work, this is done by Floyd-Warshall algorithm.

- The Floyd-Warshall algorithm exploits a relationship between path p and shortest paths from i to j with all intermediate vertices in the set $\{1, 2, \dots, k-1\}$. The relationship depends on whether or not k is an intermediate vertex of path p .
- For this algorithm, We assume that the input is represented by a weight matrix $W = (w_{ij})_{i,j \in E}$ that is defined by :
 - $w_{ij} = 0$ if $i=j$
 - $w_{ij} = w(i,j)$ if $i \neq j$ and $(i,j) \in E$

- $w_{ij} = \infty$ if $i \neq j$ and (i,j) not in E

The Floyd-Warshall Algorithm can solve the shortest path problem in $O(n^3)$ time where n = number of nodes or vertices in the graph. The steps used in the present work for computing shortest path using Floyd-Warshall algorithm is as given below:

- Let the vertices in a graph be numbered from 1 ... n.
- Consider the subset $\{1,2,\dots, k\}$ of these n vertices.
- Consider finding the shortest path from vertex i to vertex j that uses vertices in the set $\{1,2,\dots,k\}$ only. Then, there are two possible situations:
 - k is an intermediate vertex on the shortest path.
 - k is not an intermediate vertex on the shortest path.
- Let $d_{ij}(k)$ denote the weight of the shortest path from i to j such that all intermediate vertices are contained in the set $\{1,2,\dots,k\}$. we decompose p into $i \rightarrow k \rightarrow j$.
- If the vertex k is not an intermediate vertex on p , then $d_{ij}(k) = d_{ij}(k-1)$.
- If the vertex k is an intermediate vertex on p , then $d_{ij}(k) = d_{ik}(k-1) + d_{kj}(k-1)$
 - In either case, the sub-paths contain nodes from $\{1,\dots,k-1\}$. Therefore, we can conclude that $d_{ij}(k) = \min\{d_{ij}(k-1), d_{ik}(k-1) + d_{kj}(k-1)\}$ and
 - when $k=0$, then $d_{ij}(0) = w_{ij}$
 - when $k>0$, then $d_{ij}(k) = \min\{d_{ij}(k-1), d_{ik}(k-1) + d_{kj}(k-1)\}$

The shortest path matrix after applying Floyd-Warshall algorithm for 3 Tap IIR Filter is given in Figure 4.

Columns 1 through 12

1	1	2	1	2	3	1	2	3	2	1	0
0	1	2	1	2	3	1	2	3	2	1	0
0	0	2	1	2	3	1	2	3	2	1	0
0	0	2	1	2	3	1	2	3	2	1	0
0	0	0	1	2	3	1	2	3	2	1	0
0	0	0	1	2	3	1	2	3	2	1	0
Inf	Inf	Inf	Inf	Inf	Inf	Inf	Inf	Inf	Inf	0	0
Inf	Inf	Inf	Inf	Inf	Inf	Inf	Inf	Inf	Inf	0	0
Inf	Inf	Inf	Inf	Inf	Inf	Inf	Inf	Inf	Inf	0	0
Inf	Inf	Inf	Inf	Inf	Inf	Inf	Inf	Inf	Inf	0	0
Inf	Inf	Inf	Inf	Inf	Inf	Inf	Inf	Inf	Inf	0	0
Inf	Inf	Inf	Inf	Inf	Inf	Inf	Inf	Inf	Inf	0	0

Figure 4. Matrix for the generated shortest path

These critical path and shortest path computations are further used in Retiming technique to obtain the retimed circuit. This is applicable for any kind of DSP filter block set.

I a): Retiming[7][10][11] using Register Minimization Technique: Retiming can be performed majorly using three techniques:

- Cutset method
- Clock period minimization method
- Register minimization method

In Retiming using Register minimization[7], we can obtain the Digital filter that uses minimum number of registers and satisfying the clock period constraints. Here forward splitting or Register Sharing [11] is used. If the node has several output edges carrying the same signal, the number of registers required to implement these edges is the maximum number of

registers on any one of the edges. Consider Figure 5. The maximum number of registers required in 5(a) is 6 whereas after register sharing, this gets reduced to 3 as shown in Fig 5(b).

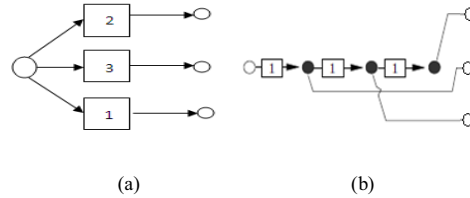


Figure 5. (a)Graph before Register Sharing (b)Graph after Register Sharing

The number of registers needed to construct this output edges(e) in retimed graph W_r and the total cost is :

$$R_v = \text{Max}(W_r(e)) \text{ and Cost} = \sum R_v$$

The cost is WRT:

- Fanout Constraints: $R_v \geq W_r(e)$ for all V and all edges $V \xrightarrow{e}$ any other vertex
- Fesibility constraints: $r(U) - r(V) \leq w(e)$ for every edge $U \xrightarrow{e} V$.
- Clock period constraints: $r(U) - r(V) \leq W(U,V) - 1$ for all vertices such that $D(U,V) > c$ where c is the clock period.

This method makes use of gadgets[Parhi] to represent the nodes with multiple edges. The register minimization retiming can be modeled as linear programming problem. A dummy node with zero computation time will be introduced in this. The weight of the edge e_i is defined to be $w(e_i) = w_{\max} - w(e_i)$ where $w_{\max} = \max(w(e_i))$ where $1 \leq i \leq K$ where k is the number of edges available. Also β parameter is used which is the breadth associated to model the memory required by edge e_i . The breadth of each edge is inverse of k . A binary search is performed for clock period and below is the procedure used while performing retiming using register minimisation.

- Use the Gadget model of the graph to compute the cost function.
- Calculate S' by using shortest path Floyd-Warshall algorithm. The Matrix obtained for 3 Tap IIR filter is:

$S' =$

22	23	49	25	51	77	25	51	77	49	-1	23	75		
-1	22	48	24	50	76	24	50	76	48	-2	22	74		
-2	-1	47	23	49	75	23	49	75	47	-3	21	73		
-3	-2	46	22	48	74	22	48	74	46	-4	20	50		
-4	-3	-2	21	47	73	21	47	73	45	-5	19	24		
-4	-3	-2	21	47	73	21	47	73	45	-5	19	-2		
Inf	Inf	Inf	Inf	Inf	Inf	Inf	Inf	Inf	Inf	-3	-2	50		
Inf	Inf	Inf	Inf	Inf	Inf	Inf	Inf	Inf	Inf	-2	-4	24		
Inf	Inf	Inf	Inf	Inf	Inf	Inf	Inf	Inf	Inf	-2	-4	-3	-2	
Inf	Inf	Inf	Inf	Inf	Inf	Inf	Inf	Inf	Inf	-2	-1	75		
Inf	Inf	Inf	Inf	Inf	Inf	Inf	Inf	Inf	Inf	Inf	Inf	77		
Inf	Inf	Inf	Inf	Inf	Inf	Inf	Inf	Inf	Inf	Inf	Inf	-1	Inf	76
Na	Na	Na	Na	Na	Na	Na	Na	Na	Na	Na	Na	Na	Na	Na

Figure 6. Matrix for S' using Floyd Warshall Algorithm.

- Compute D(U,V) and W(U,V) matrices from the original graph and S' matrix. For considered example we have W and D as below:

w =										d =														
0	1	2	1	2	3	1	2	3	2	0	1	3	1	4	4	3	3	3	3	4	2	4	3	
0	0	2	1	2	3	1	2	3	2	0	1	3	2	1	5	4	4	4	4	4	5	3	5	4
0	0	0	1	2	3	1	2	3	2	0	1	3	3	2	1	5	5	5	5	5	6	4	6	5
0	0	2	0	2	3	1	2	3	2	0	1	2	4	3	7	2	6	6	6	6	7	5	7	2
0	0	0	1	0	3	1	2	3	2	0	1	1	5	4	3	7	2	7	7	7	7	8	6	8
0	0	0	1	2	0	1	2	3	2	0	1	0	5	4	3	7	7	2	7	7	7	8	6	8
Inf	Inf	Inf	Inf	Inf	Inf	0	Inf	Inf	Inf	0	2	2	N/a	N/a	N/a	N/a	N/a	2	N/a	N/a	N/a	4	3	2
Inf	Inf	Inf	Inf	Inf	Inf	0	Inf	0	0	0	1	1	N/a	N/a	N/a	N/a	N/a	N/a	2	N/a	3	5	4	2
Inf	Inf	Inf	Inf	Inf	Inf	Inf	Inf	0	0	0	0	0	N/a	N/a	N/a	N/a	N/a	N/a	2	3	5	4	2	
Inf	Inf	Inf	Inf	Inf	Inf	Inf	Inf	0	0	0	3	3	N/a	N/a	N/a	N/a	N/a	N/a	1	3	2	3	3	
Inf	Inf	Inf	Inf	Inf	Inf	Inf	Inf	0	Inf	3	3	3	N/a	N/a	N/a	N/a	N/a	N/a	1	N/a	1	N/a	1	
Inf	Inf	Inf	Inf	Inf	Inf	Inf	Inf	0	0	3	3	3	N/a	N/a	N/a	N/a	N/a	N/a	2	1	2	1	2	
N/a	N/a	N/a	N/a	N/a	N/a	N/a	N/a	N/a	N/a	N/a	N/a	N/a	N/a	N/a	N/a	N/a	N/a	N/a	N/a	N/a	N/a	N/a	N/a	0

Figure 7. D and W Matrices for 3 Tap IIR filter

- Perform LP formulation such that the cost function gets minimised which is subjected to feasibility and clock period constraints.
- This LP problem is solved to obtain the Retiming solution which minimizes the number of registers by satisfying the clock period.

It is observed that retiming using register minimisation reduces the register count and also increasing the clock frequency. The Circuit of 3 IIR Tap filter as given in Figure 1 gets modified as shown in Figure 8(a) after performing only clock period minimization using retiming. The circuit after Retiming using register minimization is as shown in Figure 8(b). It is seen that the number of delay elements gets reduced from 6 to 4 when retiming using Register minimization is performed. Retiming using register minimization transform also reduces the clock period from 6 units (Without any retiming) to 4 units.

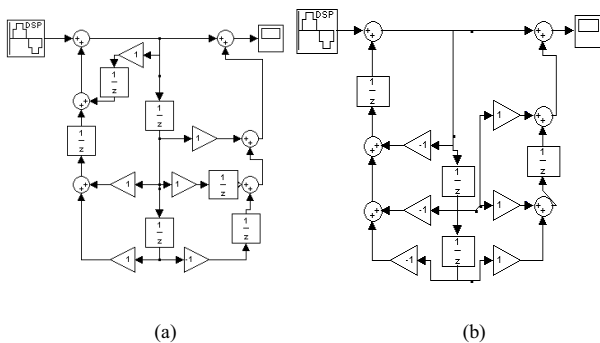


Figure 8. (a)Circuit after Clock period minimisation Retiming (b)circuit after Register minimisation retiming

For all the computations, the delay for adder is assumed as 1unit and for multiplier is assumed as 2 units. After performing retiming using register minimisation the output is taken in the form of matrices and this is used to generate synthesisable HDL code. Automation in this work is done using pearl script. This speeds up the design and testing time of Digital filters.

The further speed optimization is obtained for Digital filters by integrating Kogge-stone parallel prefix Ling adder in the Digital filters. From our earlier work [12]it is seen that

Kogge-Stone parallel prefix adder performs much better in terms of power consumption and speed when compared to other parallel prefix adders. Hence this is chosen as adder for all the filter implementations.

II a)Kogge-Stone Ling adder: Parallel-prefix adders offer a highly efficient solution in terms of speed to the binary addition problem and are well-suited for VLSI implementations. The delay[quote your ref] for all the implementations is compared by writing and simulating parallel prefix adder RTLs. It is found that the delay in the Kogge-Stone Ling adder is much less when compared to the other adders this is because the number of logic levels gets reduced in this. Hence this integrated with the retimed filter block which further increases the speed of the filter. Carry Look Adder (CLA) equations rely on adjacent pair bits (a_i, b_i) and (a_{i-1}, b_{i-1}). Along with bit generate and bit propagate, the half sum bit is present in Ling adder which is given by

$$d_i = a_i \oplus b_i$$

Instead of utilizing traditional carries, a new type of carry, known as Ling carry is produced where the i^{th} Ling carry in [12] is defined to be:

$$c_i = H_i \cdot p_i \quad \text{Where,}$$

$$H_i = c_i + c_{i-1}$$

$$\Rightarrow H_i = g_i + g_{i-1} + p_{i-1} \cdot g_{i-2} + \dots + p_{i-1} \cdot p_{i-2} \cdot p_{i-3} \cdot \dots p_1 g_0$$

Consider an example of C_4 and H_4 . If we assume that all input gates have only two inputs, we can see that calculation of C_4 requires 5 logic levels, whereas that for H_4 requires only four. The sum bit, when calculated by using traditional carry, is given to be

$s_i = d_i \oplus c_{i-1}$ which again can be minimised using Ling adders as:

$$s_i = \bar{H}_{i-1} \cdot d_i + H_{i-1}(d_i \oplus p_{i-1})$$

Above Equation can be implemented using a multiplexer with H_{i-1} as the select line, which selects either d_i or $(d_i \oplus p_{i-1})$. No extra delay is added by Ling carries to compute the sum since the delay generated by the XOR gate is almost equal to that generated by the multiplexer and that the time taken to compute the inputs to the multiplexer is lesser than that taken to compute the Ling carry. Using this Ling methodology a 16 bit adder is implemented using the Kogge-Stone tree and then that block is utilized to develop 32 and 64 bit adders. Cell other than Gray and Black cell that is used in Ling Adders are shown in Figure 9 and 10.

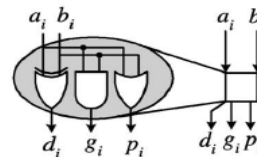


Figure 9. Bit generate, propagate and half-sum

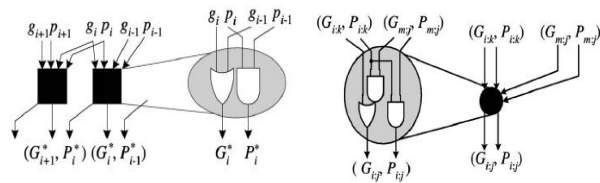


Figure 10. Ling Generate and propagate in Ling CLA

Here, the design of Ling adders by Kogge-Stone implementation is done using CMOS logic. It is seen that the combinational delay of 32 bit Ling adders by Kogge-Stone implementation is 12.492ns where as for the 32 bit ripple adder implementation, it is 14.504ns post synthesis. If the combinational delay is less, time required for computation will be less there by clock speed increase is obtained.

II. SIMULATIONS AND RESULTS

Here 3 Tap IIR filter is used as the application example. The Graphical User interface (GUI) designed for this Design optimization platform is as given in Figure 11.

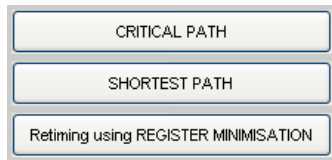


Figure 11. GUI for the designed Design optimization platform using Retiming for Register minimisation

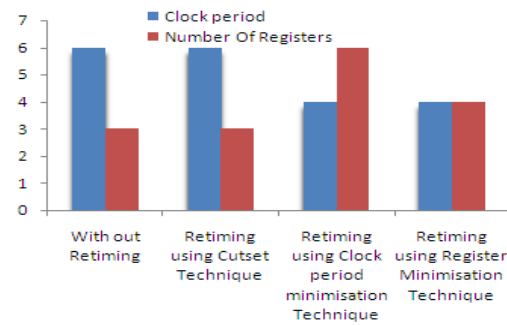


Figure 12. Graph showing Clock period and Register count using different retiming techniques for 3 Tap IIR filter

It is seen that the register minimisation method and clock period minimisation method produces the device with reduced clock period when compared to a circuit without Retiming. Due to this the operating clock frequency of the filter increases. It can be observed that the number of registers o get minimised in Register minimisation method when compared to retiming using clock period minimisation method. Hence the best compromise of speed and area is obtained in Retiming using Register minimisation method. Further the clock

frequency can be improved using Parallel prefix Kogge Stone Ling adder. After Retiming optimization, the design platform also generates the synthesisable HDL Code of the considered filter which reduces lot of designer's effort. After synthesis, the RTL Schematic of the generated 3 Tap IIR Filter is as shown in Figure 13.

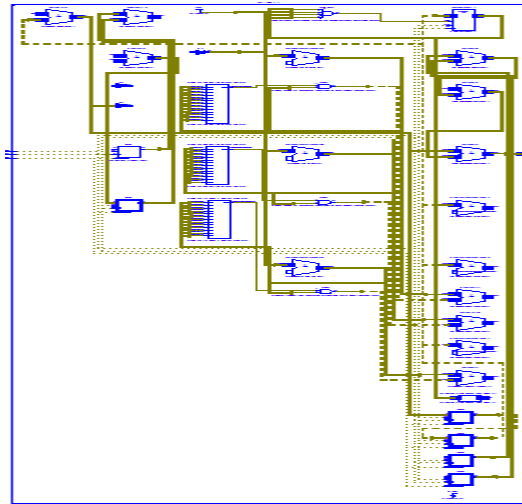


Figure 13. Schematic of the retimed 3 Tap IIR filter after synthesis of generated HDL.

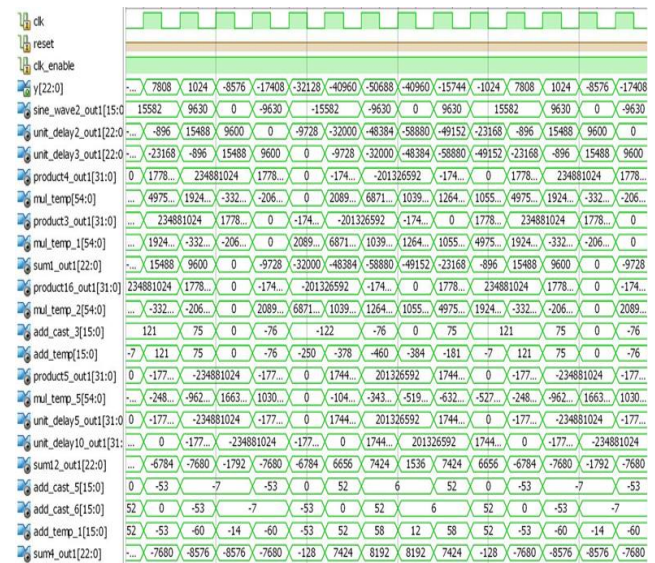


Figure 14. Simulation results of 3 Tap IIR Filter HDL.

It is seen that the post synthesis results of lattice filter matches with the matlab results with 2% error. The synthesis of lattice filter is performed on Spartan3E XC3S100E device. The number of slices used by filter after Register minimum retiming for 3 Tap IIR filter is given in Table I. The operating clock frequency is 208.69MHz for retiming using cut set and clock period minimisation method.

Architecture	with out Retiming	Retiming Using Clock period min	Retiming Using Register Min
# of Slices	100	110	107
# of LUTs	177	129	127
# of Slice FFs	97	187	180

TABLE I. DEVICE UTILIZATION SUMMARY FOR 3 TAP IIR FILTER

It is seen that the slices, lookup tables and slice flip flop counts reduce for retiming using register minimisation when compared to retiming using clock period minimisation for the same operating clock frequency. Cut set retiming method cannot be considered since operating frequency will be less in that when compared to other two methods. It is proved that register minimisation retiming is much efficient than clock period minimisation method. The different retiming methods are compared WRT the clock period (before and after retiming) and register count (before and after retiming). The results are given in Table II.

Filter Circuits	With out Retiming		Cutset Retiming		Clock period minimisation		Reg Minimisation	
	Clock period	Reg Count	Clock period	Reg Count	Clock period	Reg Count	Clock period	Reg Count
2 Tap IIR-1	30	2	25	3	25	3	25	2
2 tap IIR2-2	3	3	2	4	2	4	2	3
3 Tap IIR	6	3	6	3	4	6	4	4
4 Tap IIR	7	8	7	8	2	13	3	9
Bi-Quad-1	5	2	5	2	4	4	4	3
Bi-Quad-2	7	3	7	3	4	6	4	5
3rd order cascaded IIR	6	3	5	3	4	6	4	5
3rd order parallel IIR	8	3	8	3	5	3	4	5
8 Tap FIR	10	8	10	8	2	22	4	12
Correlator	24	4	24	4	14	5	14	4

TABLE II. CLOCK PERIOD AND REGISTER COUNT BEFORE AND AFTER REGISTER MINIMUM RETIMING FOR VARIOUS DIGITAL FILTER BLOCKS.

Along with all these, if 45 nm design library is used instead of 90nm and 180nm for the filter structures, the voltage gets minimised due to which saving in power can be obtained.

III. CONCLUSIONS

In this particular work a design optimization platform is developed for the Digital filter blocks. There are two ways in which the optimization is performed in the current work. Firstly optimization is performed using retiming using register minimisation technique. This method is chosen because for the required clock period constraints this gives the circuit with minimum number of registers. It is observed that the operating clock frequency in the digital filter can be increased to a great extent after retiming. The second type of optimization is usage of parallel prefix Ling adders instead of normal ripple adders. This also increases the operating frequency. The Design optimization platform takes the inputs from the user in the form of DFGs and matrices. This generates the synthesizable HDL. Since the entire process is automated the designer time and effort is saved if the design cycle reiterates.

- [1]. G. De Micheli, "High-Level synthesis of digital circuits", IEEE Design and Test of Computers, Oct. 1990, pp. 6-7
- [2]. A.E. Casavant, D.D. Gajski and D.J. Kuck, "Automatic design with dependence graphs", Proc. 17th ACM/IEEE Design Automation Conference, Minneapolis, June 1980, pp. 506-512
- [3]. D.D. Gajski, L. Ramachandran, "Introduction to high-level synthesis", IEEE Design and Test of Computers, Winter 1994, pp. 45-54
- [4]. L.E. Lucke, K.K. Parhi, "Data flow transformations for critical path time reduction in high level DSP synthesis", IEEE Trans. on CAD of Integrated Circuits and Systems, vol. 12, no. 7, July 1993, pp. 1064-1066
- [5]. K.K. Parhi, "Static rate optimal scheduling of iterative data-flow programs via optimum unfolding", IEEE Trans. on Computers, Vol. 40, no. 2, Feb. 1991, pp. 178-195
- [6]. S-H Huang, J.M. Rabaey, "An integrated framework for optimizing transformations", Proc. IEEE VLSI Signal Processing, San Francisco, CA, Oct. 1996
- [7]. K.K. Parhi, "High-Level algorithm and architecture transformations for DSP synthesis", Journal of VLSI signal processing, Vol 9, 1995, pp. 121-143
- [8]. H. Liao, W.W-M. Dai, "A new CMOS driver model for transient analysis and power dissipation analysis", in Low power VLSI Design Technology, G.K. Yeap, F N Najam (Eds), London, UK: World Scientific, 1996, pp 47-63
- [9]. A.P. Chandrakasan, M. Potkonjak, R. Mehra, J.M. Rabaey and R.W. Broderson, "Optimizing power using transformation", IEEE Transactions On Computer Aided Design Of Integrated Circuits and Systems, Vol. 14, no. 1, Jan 1995, pp. 12-31
- [10]. Jie-Hong R. Jiang and Robert K. Brayton. "Retiming and Resynthesis: A Complexity Perspective" IEEE trans. on CAD for integrated circuit and system, vol. 25, no. 12, 2006
- [11]. C. Leiserson, F. Rose, and J. Saxe, "Optimizing synchronous circuitry by retiming," in *3rd Caltech Conf. VLSI*, Pasadena, CA, 1983, pp. 87-116.
- [12]. Giorgos Dimitrakopoulos and Dimitris Nikolos, "High-Speed Parallel-Prefix VLSI Ling Adders", IEEE transactions on computers, Vol. 54, No. 2, February 2005
- [13]. Akansha Baliga, Deepa Yagain, "Design of High Speed Adders Using CMOS and Transmission Gates in Submicron Technology: A Comparative Study," *icetec*, pp. 284-289, 2011 Fourth International Conference on Emerging Trends in Engineering & Technology, 2011
- [14]. M. N. Mneimneh, K. A. Sakallah, and J. Moondanos, "Preserving synchronizing sequences of sequential circuits after retiming," in Proc. Asia and South Pacific Des. Autom. Conf., Jan. 2004, pp. 579-584
- [15]. S. Bomm, N. O'Neill, and M. Ciesielski. "Retiming-based factorization for sequential logic optimization", *ACM TODAES*, Vol. 5(3), July 2000, pp. 373-398.

Cloud Security Tactics: Virtualization and the VMM

Panagiotis Kalagiakos
p_kalagiakos@yahoo.com

Margarita Bora
margarita.bora@yahoo.com

Department of Information Technology
Hellenic American University
156 Hanover Street, Manchester, NH 03103, USA

Abstract— Cloud Computing is acclaimed to be the new paradigm that will reform organizations and pave the way to new and better business practices. Its abundant features captivated the interest of many individuals and organizations, either as customers or providers. The market, emboldened by optimistic reports regarding the future of cloud computing, demands further innovative services and features.

The migration to cloud computing, however, is deterred by the issue of security. Especially, in virtual environments, security is a major concern, as multi-tenancy may facilitate cyber attacks at a massive scale. The subversion of a system implicates multiple customers, augmenting the potential impact. Thus, researchers have concentrated their efforts in designing architectures and techniques that will endorse security in virtualized environments of the cloud. In this paper, we present the latest proposed techniques which aim to enhance security and revolve directly or indirectly around the most important component of the virtual environment, the Virtual Machine Monitor.

Keywords—cloud computing, security, virtual machine monitor, hypervisor.

I. INTRODUCTION

A new paradigm that has taken the business world by storm is cloud computing. The incessant innovations in technology prompted the rapid deployment of cloud computing systems in broad areas of business. According to surveys [1], [2], more than 60% of organizations and individuals are involved, in some way, with a type of cloud service. Even the federal government's agencies of the United States have inaugurated the adoption of cloud services as they comply with the "Cloud First policy" [3]. Cloud computing is heralded as the new paradigm intent to reform organizations and the way they conduct business. The migration to cloud computing, however, is deterred by the issue of security.

II. CLOUD COMPUTING SECURITY

The National Institute of Standards and Technology (NIST) has provided its own definition of cloud computing, which is highly regarded. It constitutes a

"model for enabling ubiquitous, convenient, on-demand network access to a shared pool of configurable computing resources (e.g., networks, servers, storage, applications, and services) that can be rapidly provisioned and released with minimal management effort or service provider interaction".

The cloud is distinguished by scalability, multi-tenancy, economies of scale, ubiquity and on-demand service [4]. It enables the offering of many products as a Service, rendering it highly attractive for the organizations.

However, there still remains great uncertainty and lack of information, regarding cloud computing. Many services have already migrated to the cloud and the customers are not even aware [5] that they are implementing them. This is expected as the migration occurs without the proper education or cogitation [1] and without the appropriate groundwork [7]. Due to the inordinate enthusiasm for the new trend, cloud sprawl is now a potential risk [2]. The flexibility of cloud computing further entangles issues as Infrastructure-as-a-Service (IaaS), Software-as-a-Service (SaaS) and Platform-as-a-Service (PaaS) providers collaborate and customers become dependent on more than one [1].

While the reduction of cost remains the primary reason to encourage the migration to the cloud, security is the one that deters it, due to the rising concern about the proliferation of potential risks and attacks [1]. Cyber crime is an undisputed reality that causes approximately 114 billion dollars damages each year [6]. The cloud is progressively been transformed into a magnet of potential attackers, since the subversion of such a system will compromise a multitude of targets/customers at the same time. In 2010-2011, according to a survey conducted by Trend Micro [5], more than 40% of the interviewees had encountered an issue concerning security in their cloud.

The result of the increasing risks is that a large number of customers were compelled to discontinue their cloud service [2]. Nevertheless, it is very surprising that while security is a primary concern [2], [5], [3], [7], many customers do not undertake additional measures to fortify it and almost 52% of them take none [1], [7].

Security is also an issue regarding virtualization as it is an important component of cloud computing. It is estimated that virtualization is being adopted in a very fast pace [5] as more and more organizations wish to reap its benefits. More than 50% of today cloud customers are employing virtualized services [8]. The main reason is that virtualization dissociates the hardware from the operating system (OS), empowering multiple OSs of multiple customers to be run in the same hardware platform. Customers can provision, de-provision and relocate Virtual Machines (VMs) at will without great effort and to the great benefit of the cloud providers.

The Virtual Machine Monitor (VMM), also called hypervisor, is the software which permits multiple guest VMs to run concomitantly at the same server. Moreover, it is primarily responsible of the management of the VMs, along with the available resources. It is widely acknowledged that the VMM is a crucial component that if infiltrated in any way by malicious attackers, the outcome will not just be the subversion of one guest VM/customer, but it may cause a ripple effect at a massive scale. Due to the responsibilities and the functions of the VMM, its subversion can provide access not only to the underlying resources, but the guest VMs as well. Researchers had demonstrated that it is possible to deploy guest VMs in the same environment as their target and then an attack against other VMs or the VMM would be feasible [9]. Potential threats regarding the VMM are depicted in [10] including VM escape and Hyperjacking (e.g. BluePill [11], SubVirt [12] and Vitriol [13]).

VMMs have been considered more secure than their equivalent OSs due to their smaller size. However, this argument is discarded, due to the size of the code base of today's VMMs [14] [15] [16] [17] [18], which in many cases are estimated to surpass the 200K source lines of code (SLOC) [14]. As they become more and more complex and multifarious, the possibility of vulnerabilities that can be exploited increases proportionally. That can be inferred by examining the National Vulnerability Database [19] of NIST, where it is evident that commodity VMMs have manifested a significant number of vulnerabilities that can potentially harm its users. The most popular Type I VMMs are VMware ESXi [20], Xen [21], BitVisor [22] and Microsoft Hyper-V [23]. Type II VMMs are Oracle Virtualbox [24] and VMware Workstation [25].

The Virtual Machine Monitors and their security attract the attention of many researchers as compromising the hypervisor is their primary reason that prevents organizations from deploying VMs in the cloud [7]. There are many techniques that are employed to ensure the protection of the hypervisor and the environment in which multiple guest VMs co-reside. Downsizing, strengthening and protecting the VMM represent challenges to which several researchers have risen. The security of the individual VMs/customers and the whole architecture/provider are contingent upon the deployment of schemes and features purposefully designed to promote it. Security is the widely acknowledged number one problem that can have a great impact in the future of the cloud.

This paper depicts several architectures and mechanisms which are directly associated with the security of the virtualized environment and implicate directly or indirectly the VMM. Section II denotes methods developed to strengthen and protect the VMM. Section III enumerates protection techniques involving the minimization of the VMM and Section IV introduces several schemes that can be implemented in the cloud or become the foundation for further research. Section V depicts some additional and latest efforts in VMM security and Section VI includes some discussion over the general security of the VMM.

III. STRENGTHENING AND PROTECTING THE VMM

The integrity of the VMM is deemed crucial as the emergence in cloud computing and especially in virtualized environments continues to grow in a rapid pace. The multitude and diversity of commodity VMMs guarantee functionalities that could satisfy all tastes and needs. However, the integrity of the hypervisor is an indispensable feature, so protecting and strengthening the VMM is a primary objective. In [26], several critical and necessary guidelines are depicted that would enhance the integrity and security of the VMM, along with their correlation with commodity VMMs.

A. *HyperSafe*

A new approach which strives to attain higher level of security by strengthening the VMM is HyperSafe. The integrity of the VMM in the duration of its operation is a focal point [14]. Two techniques are employed in this method, which is based on the Control-Flow Integrity mechanism (CFI) [27]. The first technique is called "non-bypassable memory lockdown" and constitutes the foundation upon which HyperSafe strives to achieve its main goal, ending a VMM with self-protection. The memory management method, illustrated in HyperSafe, is paging, as the utilized Intel x86 architecture favors it. This technique endeavors to protect memory pages and their attributes from any kind of malicious modification. It utilizes the Write-Protect (WP) bit, provided by the hardware, to alternately activate and deactivate write protection of the memory pages in case of authorized updates [14].

The second technique, called "restricted pointer indexing", is a supplement of the first, used to encompass the control data to its protection umbrella, along with memory pages. It is affiliated with the CFI mechanism. Control data in the form of return addresses and function pointers, are accumulated in target tables and then are supplanted by indexes. This is imposed by the high frequency of updates that control data are subject to and their dynamic nature. Integrity is assisted by complying with the Control-Flow Graph (CFG) [14].

The risks taken into consideration include code injection attacks and exploitation of any weaknesses in the code [14]. The hardware is based on the Trusted Platform Module (TPM) of the Trusted Computing Group (TCG) [28], so it is regarded as secure in contrast to the VMM, which is considered untrusted [14]. HyperSafe can be implemented without postulating specialized additions to commodity hardware. It is developed for Type I hypervisors and has an impact on performance in the range of 5% [14]. The existing prototypes are based on the Xen and BitVisor VMMs. However, HyperSafe imposes alterations to crucial components of the VMMs [29] and thus making it more complex.

B. *CloudVisor*

CloudVisor represents a software layer in a level below the VMM. It intervenes between the VMM and the VMs, assuming the burden of securing the resources of the guests, while resource management is still in the hands of the VMM. This segregation is achieved through nested virtualization and

protects even against infected VMMs. The Trusted Computing Base (TCB) is diminished, as now the VMM and any manager VM are excluded and only the CloudVisor is contained. It has the highest privileges, in contrast to the VMM [29].

Memory protection is ensured by the CloudVisor, by managing translations from guest-physical to host-physical addresses. It examines the extended page tables and the owners of each page to ensure that updates are legit and secure. Access to the memory of VMs by the VMM or any other manager VM is not sanctioned, apart from select circumstances. Also, CloudVisor leverages TPM and virtualized based hardware. Also, the Intel Trusted eXecution Technology [30] plays an important role.

In contrast to many other methods, CloudVisor regards cloud providers as adversaries. However, in its attack model, they are not considered outright hostile, but rather their actions could give leverage to others to exploit. Nevertheless, side channel attacks, DOS or attacks using networks are not encompassed to the threats that CloudVisor is equipped to handle. The whole scheme employs commodity hardware and a VMM with slight alterations. CloudVisor cannot service more than one VMM. The prototype developed consists of the Xen VMM and the Linux and Windows functioning as guest OSs. Its size is less than 6 KLOC. Overhead can manifest depending on the circumstances and ranges from average to nominal [29].

C. HyperSentry

In the HyperSentry approach, the integrity of the VMM is the main focus, appraised by employing, independent from the VMM, software. All integrity estimates are covert, so that any malicious behavior on the VMM cannot be concealed. The employed software obtains for its duration the higher privilege and the subsequent outcome cannot be manipulated. This software along with hardware and firmware comprise the TCB of the HyperSentry. A System Management Interrupt caused by an out-of-band channel activates HyperSentry, in System Management Mode (SMM). The SMM guarantees HyperSentry's data and code security [18]. However, researchers consider its employment as advantageous as it is limiting, because SMM can also facilitate attacks [15].

Also, HyperSentry boasts in-VMM's-context integrity assessments. The validity of the output is enforced by supplying the SMRAM with keys and then by locking it. The SMI Handler has been already copied to the SMRAM and along with the Measurement Agent, consolidated in the VMM, compose the software portion of the HyperSentry. Lastly, the validity of the output can be forwarded to external parties [18].

In its attack model, the physical locations employed to house the hardware are deemed well secured against attacks. Scrubbing attacks are considered the major threat. HyperSentry utilizes commodity hardware and leverages the TPM mechanism. The prototype was tested on the Xen VMM and for the out-of-band channel, the Intelligent Platform Management Interface was employed. HyperSentry causes slight overhead that is dependent on the frequency of its invocation [18].

Threats to the system are catalogued using the HyperSentry, but it does not obviate them. This is also facilitated by the timing of invocations, if they are too far apart [16]. HyperSentry is viewed as an approach that could be easily consolidated with others to endorse the security of the virtualized environment in the cloud [18].

IV. DOWNSIZING THE VMM

Instead of providing more functions to the VMM and other embellishments, other approaches encourage the diminution of its available attack surface and/or the restriction of its abundant features.

A. Secure MMU and H-SVM

The H-SVM, Hardware Assisted Virtual Machine, scheme promotes the restriction of the function and role of the VMM. It is founded on the "Secure MMU" mechanism [31], [32]. It supports the disjunction of memory isolation from memory management, both responsibilities of the VMM. The VMM retains the capability to consign memory at the guest VMs. However, the nested page table of each VM, created from the translation of the guest-physical to host-physical address, can only be modified under the supervision of an additional hardware component instead of the VMM. This hardware processor affirms any updates to the nested page tables. They are compiled to protected memory and attainable only by H-SVM. The goal is to ensure memory isolation, since already assigned physical pages cannot be accessed by another party [32].

Additional functions include the deletion of any information from any previously assigned pages, page swapping and a degree of restricted page sharing. It exhibits many similarities to CloudVisor, which however is not based on hardware, but rather on software [29].

The H-SVM mechanism assumes that the hardware used by the cloud provider is trustworthy, due to its susceptibility to this kind of attacks. The TCB is significantly minimized and it excludes the VMM, whereas incorporates the hardware processor. This whole architecture entails certain changes to the existing hardware processors, but minimal interference to the OSs and the VMM. The processors are altered in a way resembling the TPM to facilitate communication with the provider, but in the H-SVM scheme, the private key is consolidated in the processors. Also, an alternative realization of this architecture comprises of microcode routines, upon which this method was tested, modifying the Xen VMM. Performance can deteriorate due to page mapping and page swapping. Although in most cases performance overhead remains insignificant [32].

B. NoHype

Unlike to any other approaches regarding the VMM, a new approach removes it entirely from the equation. The attack surface that can be exploited by hackers is eliminated to the NoHype architecture. It regards VM exits as a potential liability that can negatively impact security and result in code injection and the subversion of the VMM [16].

The attack model used to illustrate NoHype reckons cloud providers as trustworthy. The physical locations employed to house the hardware are deemed well secured against attacks. However, the guest OSs are viewed as possible security threats.

NoHype necessitates a slight alteration of guest OS kernels, which will be provided by the cloud provider to be used in booting the VMs. While booting occurs normally, a temporary VMM is used to handle configuration data of the system and customer code is prohibited from execution while the temporary VMM is online. The customers are primarily accountable of the security of their VMs. Also, the management software used by the cloud providers is considered an issue worthy of further consideration, but for the moment trustworthy [16].

Each VM is pre-assigned its own processor core and physical memory. It is responsible for its own allocated resources. This is greatly facilitated by the cloud computing architecture as a customer commissions resources at the beginning. The local Advanced Programmable Interrupt Controller (APIC) is used by the guest VMs without any intervention. Also, VMs command their own committed I/O devices, which support virtualization. Hardware is employed as the medium to impose memory isolation [16].

However, it still leaves a window of opportunity, in the form of management software, to be exploited [15] and relinquish cardinal aspects of virtualization in the cloud [29]. The whole NoHype architecture is based upon commodity hardware. In the designed prototype, the temporary VMM is Xen 4.0 and the guest OS is Linux 2.6. Evaluations depict improvement in performance by 1% [16]. Nevertheless, there is still skepticism on whether NoHype can be properly implemented leveraging present processors [17] or advancements in that area must be first realized.

C. HyperWall

Another architecture designed by the researchers of NoHype is HyperWall, which is viewed as a more effective method in endorsing security in the virtualized environment. The amalgamation of HyperWall with other models (e.g. HyperSafe) could eventuate in a more secure architecture. This approach considers that the VMM is a potential threat and consequently is banned from the TCB. HyperWall advocates the use of the “Confidentiality and Integrity Protection” (CIP) tables. This important characteristic of HyperWall, guarantees that the memory of the VMs is protected according to individual preferences, stated by customers. This protection from either the VMM or DMA is imposed with the assistance of the hardware, which is considered a vital component in a secure architecture [15].

Similar to other models, the VMMs retain resource management. HyperWall employs mechanisms (e.g. hash and trust measurements) that assess the state of the VMs and whether the system is trustworthy or not. In the attack model employed, hardware is deemed secure. The customers are accountable for the protection of their OS and other applications in the VMs. Nevertheless, some types of attacks (e.g. DoS, covert and side channel) are not reckoned with. This model utilizes commodity VMMs, but hardware needs to be

subjected to minor alterations. Some overhead is encountered due to the novelty of the approach, but further testing would be undertaken in future research [15]. HyperWall advertises more advanced functions, compared to other efforts.

V. ADDITIONAL TECHNIQUES

The intricacy and multifariousness of cloud computing deters the application of many existing techniques which are employed outside the cloud. Nevertheless, a number of techniques can be implemented in the cloud, or at least present a foundation for other more advanced techniques to be developed, while taking account of the particular characteristics of the cloud like multi-tenancy.

A. NOVA

A model that follows a different approach to virtualized environments is the Nova architecture. Built in a novel way, Nova strives to provide a new solution to the security issues in virtualization, by reducing the available attack surface. In the majority of approaches, the terms of the Virtual Machine Monitor (VMM) and the Hypervisor are used interchangeably. The Nova model, however, uses those terms to denote the user level and the privileged level hypervisor respectively. At the same time, it negates any usage of privileged domains.

B. sHype

The sHype architecture, developed by IBM Research, is designed to endorse isolation and security. It is founded upon the TPM to assess the integrity of the hypervisor and the VMs. Each VM is consigned its own TPM instance, which are in software form. The sHype model is comprised of the policy manager, the access control module (ACM), the secure services VM and the mediation hooks. A security policy is imposed in order to compel isolation of the VMs. It is governed by the policy manager, which dictates how the ACM conducts resource management between the VMs and generally all pivotal functions. The ACM is an integral part of the hypervisor. The sHype model was realized in different commodity hypervisors including Xen.

C. TrustVisor

TrustVisor is a small VMM designed to conduce integrity and security of the data, code and its execution. It also follows the principle of the smaller is more secure, regarding its size, which does not exceeds 7 KLOC. Furthermore, it can furnish outside verifiers with evidence of security. TrustVisor employs TPM hardware and its own TrustVisor Root of Trust Module (TRTM), which is associated with the consolidated in the TrustVisor μ TPM software. Overhead has been estimated around 7% while running TrustVisor. Although the TCB is minimized, it may require protracted development and testing as specialized to this scheme applications are incumbent.

D. SecVisor

Integrity of the code is the main focus of the SecVisor VMM. User policies define the code that will be executed, so that illegal alterations of the OS kernel or any unauthorized

code are prohibited. Its smaller TCB size guarantees limited attack surface which could be potentially exploited. Regarding memory protection, SecVisor leverages hardware and more specifically the MMU of the CPU. SecVisor is designed to preclude code injection attacks by securing the kernel. However, its attack model does not include control-flow threats. It requires hardware that supports virtualization and has been implemented in the case of one CPU. SecVisor considers that it is secure in its own right and focuses on guarantying the OS kernel integrity. The conviction that SecVisor could be consolidated with other security schemes to provide an overall level of security is deemed feasible [14].

VI. DISCUSSION

The above research efforts reflect many of the latest innovations and schemes that researchers have produced in order to provide additional security to the VMs and the cloud. Those schemes are focused in security and the looming threat of cybercrime.

Security is the major issue that can be the cause of perturbation and lack of trust in the part of customers, who can envision their confidential and sensitive data compromised and their systems breached. Their resulting losses could range depending on their commitment to the cloud, but it would undoubtedly be detrimental to the reputation not only of the cloud provider but to the cloud in its entirety.

Several additional features or overall architectures have ensued due to the different security concerns, in order to fortify the structure of the virtualized environment against cyber crime. Strengthening and protecting the VMM is a popular area of research and patronized by several researchers (e.g. [14], [18], [29]). These techniques include protection either of the VMs or the VMM and revolve around the VMM.

It is understandable that cloud providers and VMM designers would endeavor to make their products more appealing to customers by interminably adding more and better attributes. The resulting VMMs have more impressive capabilities and contribute to a better cloud virtualized environment. However, all these functions that are an inducement for the potential customer are becoming the system's Achilles heel by endangering its security. Many consider that all these extras should be segregated from the core of the VMM, which would be only comprised of its essential components.

Downsizing the VMM is a method that has various proponents, who implement it in various degrees (e.g. [31], [32]). The subsequent architecture may or may not have the VMM in a prominent position. The decrement of the TCB is an approach that captivates many researchers and is pursued with either software or hardware means. The ultimate goal is to decrease the attack surface and thus reduce the possibilities of a cyber attack.

In contrast, some researchers (e.g. [29]) consider putting a software component or a nested VMM beneath the VMM to enforce security. The additional software would be smaller in size code and thus more protected against malicious attacks.

This solution, however, is repudiated by many [14] [15], due to the subsequent issue of securing that layer.

The hardware can be leveraged to play an important role to security and it is considered trustworthy. Many conform to the TPM specifications [14], [18], [29] or they employ resembling techniques. Although, it is found that TPM focuses primarily on the loading stage and not on the duration of the operation of the VMM [14] and thus leaving an opening for exploitation. Integrity attestation in the whole duration of the operation of the VMM features prominently in many research efforts (e.g. [14], [18], [15]).

One important issue that is being addressed by the schemes above is memory management. The commodity VMM is in total and undeniable control of securing the memory. It designates specific memory pages to the VMs and dictates translations from guest-physical to host-physical addresses. If the VMM is subverted, an attacker could acquire access to the memory of the guest VMs and proceed to a number of subsequent malicious actions. Memory is one of the vulnerabilities of virtualization because it is shared, so it rivets most researchers, who devote considerable efforts in safeguarding it.

Researchers (e.g. [15], [18], [29], [32]) capitalize on the MMU and the IOMMU of the CPU in order to promote memory protection. Hardware that supports virtualization takes also a prominent role in the battle against cyber crime.

All methods mentioned rely in the assumption that the physical locations in which all the equipment is hosted are secure by the cloud providers. However, a divergence is observed over the intentions of the cloud provider/administrator who either is considered trustworthy or malicious.

The diverse proposals can be considered as unilateral in addressing the security issues, as their respective researchers have acknowledged. The contrived assumptions do not correlate to the overall reality of the cloud virtualized environments. One solution widely proposed is the capitalization of all researches. The amalgamation of manifold techniques (e.g. [15], [18], [29]), which revolve about distinct security issues, could eventuate to a more comprehensive and secure cloud.

Researchers cautions that the provision of a VMM, in order to support a commodity VMM, could eventually become a necessity, as commodity VMMs mature. Architectures based on existing VMMs tend to be less time-consuming to develop and possibly the cost is not as excessive. The other alternative is to redesign the VMM without following the same path that could lead to the same obsolete solutions. The first research efforts using new technologies have begun to materialize.

With the recent innovations in virtualization technology, all systems could be adopting VMMs in the future, in order to advance their functionality and protection, regardless of whether they are inside or outside the cloud. So, it is incumbent to ensure a high level of security as they are and will be the front line of defense in the complex and multifaceted virtualized computing environment, in and out of the cloud.

VII. CONCLUSION

As the demand and participation into cloud computing increases, the amount of critical and sensitive data in the cloud and the VMs would eventually become, if not already, the premium target for malicious attackers. The VMM could provide opportunities to cloud users with malicious intentions to inflict damage at a massive scale, due either to its inherent vulnerabilities or other imperfections to the architecture of the system. So, researchers have focused their attention and their resources in developing security techniques that could enhance the security of the VMM and of the entire structure. The intricacies of virtualization and the complexity of the cloud require multifaceted solutions that would prove to be adept opponents against malicious adversaries and cyber crime.

REFERENCES

- [1] Ernst & Young, "Into the cloud, out of the fog: Ernst & Young's 2011 Global Information Security Survey," 2011. [Online]. Available: [http://www.ey.com/Publication/vwLUAssets/Into_the_cloud_out_of_the_fog-2011_GISS/\\$FILE/Into_the_cloud_out_of_the_fog-2011%20GISS.pdf](http://www.ey.com/Publication/vwLUAssets/Into_the_cloud_out_of_the_fog-2011_GISS/$FILE/Into_the_cloud_out_of_the_fog-2011%20GISS.pdf).
- [2] Avanade Inc, "Global Survey: Has Cloud Computing Matured?," 2011. [Online]. Available: http://www.avanade.com/Documents/Research%20and%20Insights/FY11_Cloud_Exec_Summary.pdf.
- [3] MeriTalk, "Federal Cloud Weather Report," [Online]. Available: http://www.meritalk.com/pdfs/MeriTalk_Federal_Cloud_Weather_Report.pdf.
- [4] P. Mell and T. Grance, "The NIST definition of cloud computing: Recommendations of the National Institute of Standards and Technology (SP800-145) [Report]," 2011. [Online]. Available: <http://csrc.nist.gov/publications/nistpubs/800-145/SP800-145.pdf>.
- [5] Trend Micro, "Cloud Security Survey Global Executive Summary: Corporate Marketing," 2011. [Online]. Available: http://www.trendmicro.com/cloud-content/us/pdfs/about/presentation-global-cloud-survey_exec-summary.pdf.
- [6] Symantec, "Norton Study Calculates Cost of Global Cybercrime: \$114 Billion Annually," 2011. [Online]. Available: http://www.symantec.com/about/news/release/article.jsp?prid=20110907_02.
- [7] Event Tracker, "2010 State of virtualization security survey: Current opinions, experiences and trends on the strategies and solutions for securing virtual environments," Prism Microsystems, 2010. [Online]. Available: <http://www.prismmicrosys.com/documents/VirtualizationSecuritySurvey2010.pdf>.
- [8] CDW, "From tactic to strategy: The CDW 2011 Cloud Computing Tracking Poll," 2011. [Online]. Available: <http://webobjects.cdw.com/webobjects/media/pdf/Newsroom/CDW-Cloud-Tracking-Poll-Report-0511.pdf>.
- [9] T. Ristenpart, E. Tromer, H. Shacham and S. Savage, "Hey, you, get off of my cloud: Exploring information leakage in third-party compute clouds," in Proceedings of the 16th ACM conference on Computer and communications security CCS, 2009, pp. 199-212.
- [10] B. Williams and T. Cross, "Virtualization System Security," IBM, 2010. [Online]. Available: <http://blogs.iss.net/archive/papers/VirtualizationSecurity.pdf>.
- [11] J. Rutkowska, "Subverting Vista Kernel For Fun And Profit," in Symposium on Security for Asia Network SyScan and Black Hat Briefings, 2006.
- [12] S. T. King, P. M. Chen, Y.-M. Wang, C. Verbowski, H. J. Wang and J. R. Lorch, "SubVirt: Implementing malware with virtual machines," in Proceedings of the 2006 IEEE Symposium on Security and Privacy, 2006, pp. 314-327.
- [13] D. D. Zovi, "Hardware Virtualization Rootkits," [Online]. Available: http://www.theta44.org/software/HVM_Rootkits_ddz_bh-usa-06.pdf.
- [14] Z. Wang and X. Jiang, "HyperSafe: A Lightweight Approach to Provide Lifetime Hypervisor Control-Flow," in Proceedings of the 2010 IEEE Symposium on Security and Privacy SP, 2010, pp. 380-395.
- [15] J. Szefer and R. B. Lee, "Architectural Support for Hypervisor-Secure Virtualization," in Proceedings of the seventeenth international conference on Architectural Support for Programming Languages and Operating Systems ASPLOS, 2012, pp. 437-450.
- [16] J. Szefer, E. Keller, R. B. Lee and J. Rexford, "Eliminating the hypervisor attack surface for a more secure cloud," in Proceedings of the 18th ACM conference on Computer and communications security CCS, 2011, pp. 401-412.
- [17] L. Gu, A. Vaynberg, B. Ford, Z. Shao and D. Costanzo, "CertiKOS: A Certified Kernel for Secure Cloud Computing," in Proceedings of the Second Asia-Pacific Workshop on Systems APSys, 2011.
- [18] A. M. Azab, P. Ning, Z. Wang, X. Jiang, X. Zhang and N. C. Skalsky, "HyperSentry: enabling stealthy in-context measurement of hypervisor integrity," in Proceedings of the 17th ACM conference on Computer and communications security CCS, 2010, pp. 38-49.
- [19] National Vulnerability Database, [Online]. Available: <http://nvd.nist.gov/home.cfm>.
- [20] VMware, "VMware ESXi," 2012. [Online]. Available: <http://www.vmware.com/products/vsphere-hypervisor/overview.html>.
- [21] Xen, 2012. [Online]. Available: <http://www.xen.org/>.
- [22] T. Shinagawa, H. Eiraku, K. Tanimoto, K. Omote, S. Hasegawa, T. Horie, M. Hirano, K. Kourai, Y. Oyama, E. Kawai, K. Kono, S. Chiba, Y. Shinjo and K. Kato, "BitVisor: A thin hypervisor for enforcing i/o device security," in Proceedings of the 2009 ACM SIGPLAN/SIGOPS international conference on Virtual execution environments VEE, 2009, pp. 121-130.
- [23] Microsoft, "Hyper-V," 2012. [Online]. Available: <http://www.microsoft.com/en-us/server-cloud/windows-server/hyper-v.aspx>.
- [24] Oracle, "VirtualBox," [Online]. Available: <https://www.virtualbox.org/>.
- [25] VMware, "VMware Workstation," [Online]. Available: <http://www.vmware.com/products/workstation/new.html>.
- [26] A. Vasudevan, J. M. McCune, N. Qu, L. Van Doorn and A. Perrig, "Requirements for an integrity-protected hypervisor on the x86 hardware virtualized architecture," in Proceedings of the 3rd international conference on Trust and trustworthy computing TRUST, 2010, pp. 141-165.
- [27] M. Abadi, M. Budi, Ú. Erlingsson and J. Ligatti, "Control-flow integrity principles, implementations, and applications," ACM Transactions on Information and System Security (TISSEC), vol. 13, no. 1, 2009.
- [28] Trusted Computing Group, "Trusted Platform Module," 2012. [Online]. Available: http://www.trustedcomputinggroup.org/developers/trusted_platform_module/.
- [29] F. Zhang, J. Chen, H. Chen and B. Zang, "CloudVisor: Retrofitting protection of virtual machines in multi-tenant cloud with nested virtualization," in Proceedings of the Twenty-Third ACM Symposium on Operating Systems Principles SOSP, 2011, pp. 203-216.
- [30] Intel, "Malware Reduction: Intel® Trusted Execution Technology (Intel® TXT)," [Online]. Available: <http://www.intel.com/content/www/uk/en/architecture-and-technology/trusted-execution-technology/malware-reduction-general-technology.html?wapkw=trusted+execution+technology>.
- [31] S. Jin and J. Huh, "Secure MMU: Architectural support for memory isolation among virtual machines," in IEEE/IFIP 41st International Conference on Dependable Systems and Networks Workshops, 2011, pp. 217-222.
- [32] S. Jin, J. Ahn, S. Cha and J. Huh, "Architectural support for secure virtualization under a vulnerable hypervisor," in Proceedings of the 44th Annual IEEE/ACM International Symposium on Microarchitecture (MICRO-44 '11), 2011.

Designing an active band-pass filter with tunable transversal element at 4-GHz using 0.2 μ m GaAs technology

Saeed Soleimany
Islamic Azad University
Qazvin branch
Qazvin – Iran
Saeed.soleimany@gmail.com

Mohammad Reza Salehifar
Islamic Azad University
Science and Research Branch
Tehran – Iran
Mr_salehifar@yahoo.com

Hassan Karbalaee
Shahed University
Tehran – Iran
H79_karbalaee@yahoo.com

Abstract: Nowadays, 4-GHz band-pass filters become important for several fields of applications such as Bluetooth, wireless LAN (WLAN), RFIDs, and etc. The growing requests for ultra integration of small mobile communication leads the designs to more wireless applications (wireless network, mobile, and etc). A transceiver system needs to design a filter that properly provides its requirements. The filter reported in this paper uses lumped elements to achieve a basic band-pass filter response. Active transversal elements are used to sharpen the band-pass characteristics, and to overcome the losses of MMIC lumped elements. In addition, the combination of lumped and transversal elements provide much better performance compare to a filter made of lumped elements alone, and much smaller than a filter made of transversal elements alone. A GaAs MMIC active band-pass filter structure is proposed for operation in the S-band applications and GaAs MMICs applications for wireless personal communication are described. This filter centered at 4-GHz and has 30 dB rejections at 0.7-GHz apart from pass-band edges. These sharpened edges were caused by a tuned amplifier.

Keywords: Active Band-pass filter, transversal, tuned amplifier, 0.2 μ m GaAs Technology

I. INTRODUCTION

Microwave radio electronics in recent years has been substantial progress such as prospering in telecommunication market, Improvements in the working frequencies, Processing of the microwave field based on silicon and etc. Designers of high-frequency circuits, despite these advances, still have problems such as high scale integration, system configuration capability according to defined softwares in radio systems. For both cases, the special problems of filtering should be resolved. Active filtering can be one of the solutions [1] to [5].

The main feature of active filtering is in its integration with other components on a chip. Because the active filters, unlike the passive filters, have not large size resonator elements that can prevent from integrating. In most cases of the using from active filters, we are dealing with long wavelengths. For example, GSM (second generation of mobile phones) 800MHz to 1800MHz, UMTS (third generation mobile cellular technology)

1.9GHz to 2.1GHz, Wireless personal communications and wireless local area networks (WLAN, WPAN).

In today's world, with rapid expansion and increasing technology and increasing demand for mobile phones with very low weight, RFICs, Bluetooth, wireless LAN and etc., the human need to GaAs MMIC technology increases. Regardless of the reasons why the integration of mobile communication devices mentioned, one of the main problems that are faced in the filter design, is the low inductor's quality factor. Therefore, implementation of active filters in GaAs MMIC technology in the microwave frequency range is highly desirable, taking full advantage of the small size lumped element in order to achieve a basic filter response and the consequent neutralization effect of the losses achieved by the active device.

The proposed filter, also with GaAs MMIC technology has been designed and used for applications in wireless networks and Bluetooth. This paper present the computer simulation (ADS software) of the novel bandpass filter, with a basic filter structure based on the lumped and transversal technique [6] and it includes a tuned amplifier for implementing the transversal element. This tuned amplifier in transversal element allows not only the implementation of the transversal gain of the conventional transversal filters, but also the filtering of the upper and lower frequency, reducing the number of lumped elements in the structure.

II. NOVEL CIRCUIT STRAUCTURE

High-frequency monolithic integrated filters are usually a major constraint, the need of bulky resonant elements and the operational amplifier with high gain. A band-pass filter structure with lumped elements in Figure 1 is shown.

This figure shows an Nth order Chebychev filter, using the low-pass filter, amplifier and high-pass filter. As can be seen, a band-pass filter is obtained from the connection between the two low-pass and high-pass filter. It is for this purpose, between the two filters, the active components can be used, so this structure can be use to design an active filter. In other words, lumped elements to achieve the pass filter are used and transversal active elements are used for sharpening band-pass filter characteristics and also overcome the high losses of lumped MMIC devices. It should be noted that the inductor can be implemented with transmission lines or inductors ring.

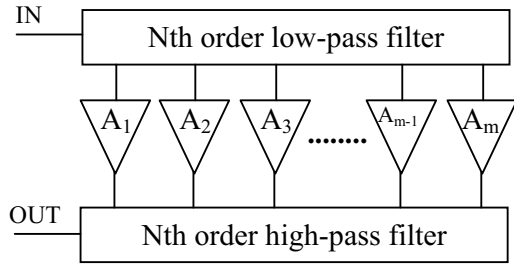


Fig 1. Conventional microwave lumped and transversal filter structure

In the proposed new filter structure, shown in Figure 2, a variable transversal element has been added. This transversal variable element, with the combined unilateral techniques and tuned amplifier is obtained and in the main signal path is placed. On the other hand, in the usual one-sided filter structure shown in Figure 1, the number of pairs of complex conjugate poles is always even, whereas, in the new structure of figure 2, this number is odd that due to the tuned amplifier. Because this tuned amplifier has a complex conjugate pole pair and a complex conjugate zero pair. This degree of freedom in design, allowing simple one-sided filter structure and compressed with Nth order.

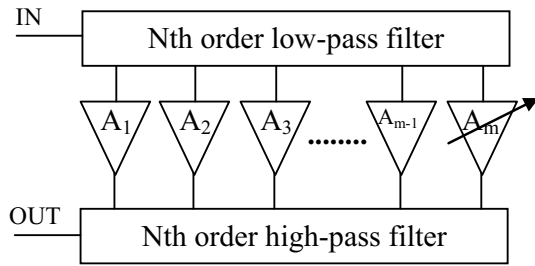


Fig 2. New microwave filter structure with a lumped and tuned transversal elements

Considering that, this amplifier is located in the main path of signal then for matching; both sides shall be terminated 50-ohm impedance and ensure stability of the amplifier will result.

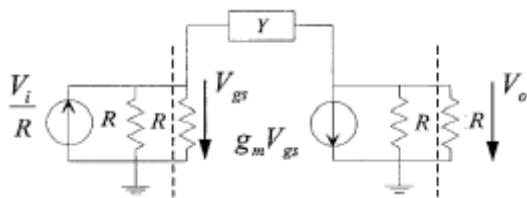


Fig 3. Ideal equivalent circuit of a tuned amplifier

Where:

$$R = 50\Omega$$

$$\gamma = SC + \frac{1}{SL}$$

In this case, according to Figure 3, the voltage transfer function of this circuit will be as follows:

$$v_o/v_i = \frac{1}{4 + Rg_m} \left[\frac{s^2 LC - sLg_m + 1}{s^2 LC + s \frac{4L}{R(4 + Rg_m)} + 1} \right] \quad (1)$$

Also, tuned amplifier S21 parameter, as:

$$s_{21} = \frac{2}{4 + Rg_m} \left[\frac{s^2 LC - sLg_m + 1}{s^2 LC + s \frac{4L}{R(4 + Rg_m)} + 1} \right] \quad (2)$$

Given that:

$$\omega_z = \omega_p = \frac{1}{\sqrt{LC}} \quad (3)$$

$$Q_z = \frac{1}{g_m} \sqrt{\frac{C}{L}} \quad (4)$$

$$Q_p = \frac{R(4 + Rg_m)}{4} \sqrt{\frac{C}{L}} \quad (5)$$

Equation 2 can be rewritten as follows.

$$s_{21} = \frac{2}{4 + Rg_m} \left[\frac{\frac{s^2}{\omega_z^2} - \frac{s}{\omega_z Q_z} + 1}{\frac{s^2}{\omega_p^2} + \frac{s}{\omega_p Q_p} + 1} \right] \quad (6)$$

In this tuned amplifier, filter frequency response is band-pass type if $Q_p/Q_z > 1$, and is notch type if $Q_p/Q_z < 1$.

The significant point is when this frequency response is combined with the Chebyshev filter that includes two low-pass and high-pass structure (form a band-pass filter), behaves like an elliptic filter. This new structure feature is an important feature in the design of filters, because it will introduce a new type of filtering characteristics which is impossible to obtain with the conventional transversal filters.

Figure 4 shows the filter design process. Figure 4(a) shows a 7th order band-pass Chebyshev that design using 7th order low-pass Chebyshev filter and 7th order high-pass Chebyshev filter. Figure 4(b) represents a filter with transversal and lumped elements that is conventional filter structure. Finally, Figure 4(c) shows a filter structure with lumped and transversal elements that is new filter structure.

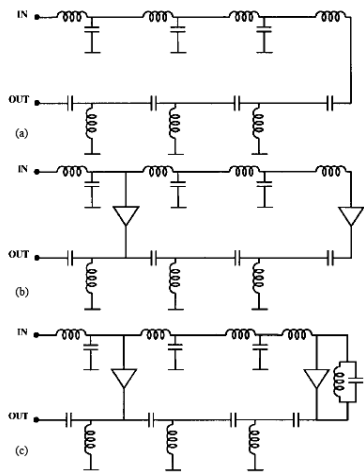


Fig 4. (a) 7th order 2dB ripple traditional Chebyshev bandpass filter, (b) Lumped and transversal filter, (c) Lumped and tuned transversal filter

An important property is that the new filter with tuned transversal element has more sharp edges of the stop band [7].

Figure 5 shows a comparison between the three filters. We present their frequency response as simulated by ADS 2010 software. The filters are designed at 7GHz center frequency.

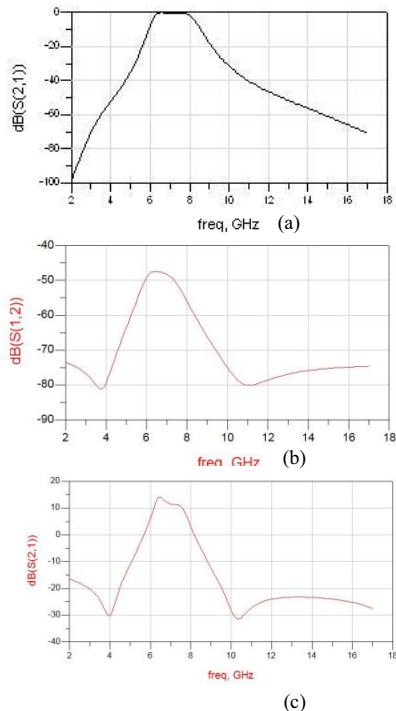


Fig 5. simulated bandpass characteristics of the 7th order (a) traditional Chebyshev filter, (b) lumped and transversal filter, (c) lumped and tuned transversal filter

Figure 5(a) shows simulation result ($S_{2,1}$ parameter) of 7th order Chebyshev conventional filter. As can be seen in figure 5(a), due to the use of passive elements, filter has not a good gain and amplitude of Band Edge Rejection also has not a good slope. Figure 5(b) represents $S_{2,1}$ parameter of 7th order Chebyshev filter with two transversal amplifier with constant gain. As can be seen, the circuit parameters are improved compared to the passive filter. Figure 5(b) shows $S_{2,1}$ parameter of 7th order lumped and tuned transversal filter. For reasons described in previous sections, this tunable transversal element, improves gain parameters and increase the band edge removing, dramatically.

III. CIRCUIT ARCHITECTURE

Figure 6 shows a Chebyshev bandpass filter with the GaAs technology. Because the amplifier gain is not desired, the two amplifier circuits, to increase the gain, are used. In figure 7, a filter amplifier circuit with the use of GaAs technology is shown. As can be seen, from a 50-ohm input and output matching is used. As noted, each amplifier provides a 180 degrees phase, therefore, the number of amplifiers, must always be paired. Figure 8 shows the complete designed circuit.

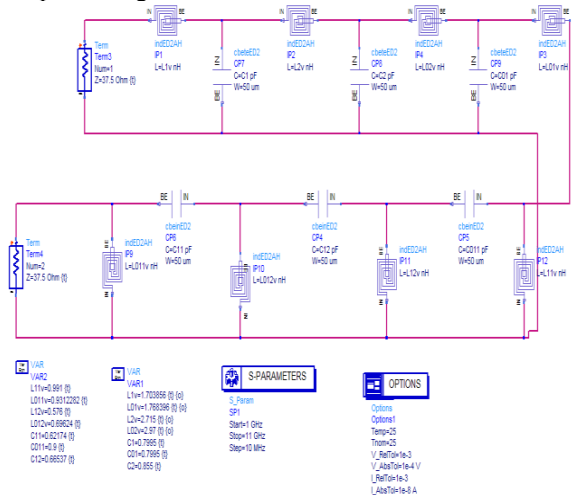


Fig 6. The bandpass filter circuit designed at the central frequency of 4GHz

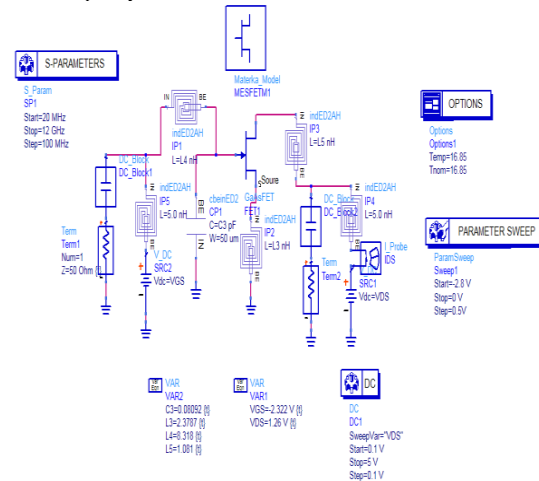


Fig 7. The designed amplifier circuit for lumped and tuned transversal Chebyshev bandpass filter

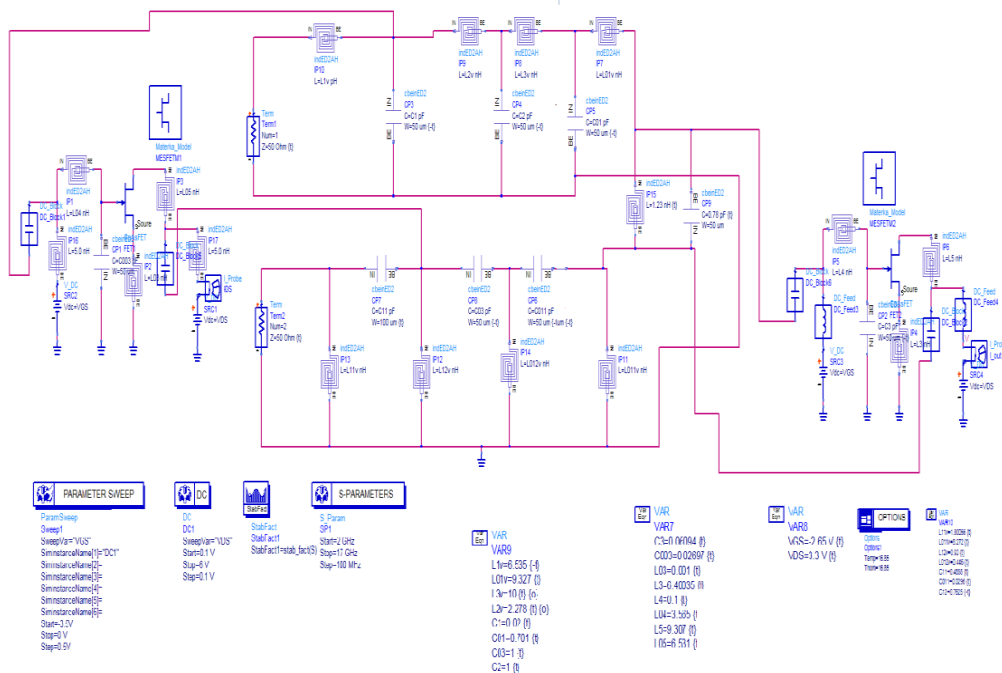


Fig 8. Circuit architecture of MMIC lumped and tuned transversal active filter using 0.2µm GaAs technology

IV. SIMULATION RESULTS

Figure 9 shows S-parameters of proposed filter. As can be seen, according to the curves of S11 and S22, the input and output matching well done. Also known as the S21 curve, in the central frequency at 4 GHz, filter gain is approximately equal to 25 dB. This filter also has the ability of adjust in the range 400 to 500 MHz.

Since the designed filter is bandpass filter, we can easily use the S21 curve to calculate quality factor (Q). First, with respect to the S21 curve, 3dB gain bandwidth got:

$$BW = 0.1 \text{ GHz}$$

$$Q = \frac{f_0}{BW} = \frac{4}{0.1} = 40$$

Power consumption in the proposed filter regardless of the inductor resistance is zero, because the filter is non-active. However, the input and output matching networks have power consumption, that its value can be achieved from value of Vds and Ids. Then circuit simulated by the ADS software, we can obtain all DC voltages and currents of the circuit.

$$P = (V_{ds} * I_{ds})$$

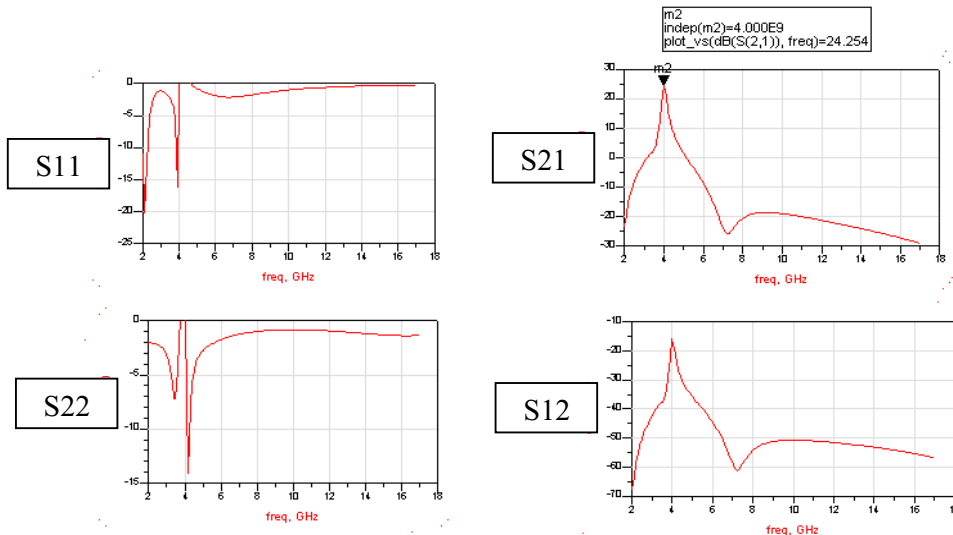


Fig 9. S-parameters Characteristics for the bandpass filter circuit with 4GHz central frequency

Power consumption in the filter, with $V_{ds} = 2.5V$ and $I_{ds} = 41$ mA at each transistor, is equal to 102.5 mW.

V. CONCLUSION

This paper presents design and simulation of an active filter that matched to 50Ω by active elements. Although, the general nature of the circuit uses the advantages of active structures, due to its simplicity of filter design and its implementation will be very easy and reliable. Also, the combination of active matching circuits, according to simulation results, indicates that we can achieve to narrow-band filtering by low order filter without repeating the pass band at higher frequencies that this low-order filter reduces designed filter size in comparison with the dimensions of high-order filters.

Simulation results note that adding two stages with active elements at both ends of passive filters greatly improves filter efficiency, among which, we can note to create the appropriate gain in the central frequency, the filter input and output matching impedance of 50 ohms, the high stability of the circuit and increase the quality factor (Q).

VI. REFERENCES

- [1] G.L.Matthael, L.Young, E.M. T. Jones, "microwave filters, impedance matching networks and coupling structures", Artech House, Dedham, mass., 1980
- [2] Stefan Anderson, Christer Svensson, "An active recursive RF filter in $0.35\mu\text{m}$ BiCMOS", Analog integrated circuits and signal processing, 44, 213–218, Springer 2005
- [3] Roberto Gomez–Garcia, Cesar Briso–Rodriguez, Mustapha Mahfoudi and Jose I. Alonso, " MMIC Tunable Transversal Bandpass Active Filterat 9–12 GHz", 11th GAAS Symposium - Munich 2003
- [4] A. Alahyari, A.Habibzadeh, M. Dousti, " Tunable Active Dual-Band Bandpass Filter Design Using MMIC Technology", International Journal of Engineering & Technology IJET-IJENS Vol: 11 No: 01, Feb 2011
- [5] U.Karacaoglu, S.Lucyszyn, I.D.Robertson and M.Guglielmi, "GaAs MMIC Active Filters for L-Band Mobile Systems", The institution of electrical engineers, 1994
- [6] M.J.schindler, Y.Tajima, "A novel MMIC active filter with lumped and transversal elements", IEEE tans. Microwave theory tech., Vol.37, pp- 2148-2153, Dec. 1989
- [7] K.W.Tam, P. Vitor, R.P.Martins, " MMIC Active Filter with Tuned Transversal Element", IEEE Transactions on circuits and systems II, Vol.45, No.5. May 1998

Determination of QoS Metrics in Wireless Sensor Networks by Using Queuing Theory

Anar Rustamov^{1,2}

¹Institute of Cybernetics, Baku, Azerbaijan

²Qafqaz University, Baku, Azerbaijan

anrustemov@qu.edu.az anar.rustemov@gmail.com

Abstract— Traffic classes in multimedia WSNs are required high bandwidth and reliable transfer through the network because of large amount of data size. However, some applications of WSNs are required high quality. In this scenario, main purpose in designing of the sensor networks and nodes becomes how to conserve energy and prolong network life. In order to increase quality of services (QoS), a measurement of probability of blocking of the arrival packets were suggested by using queuing theory. Probability of blocking gives us clear picture how system specification should be chosen so that blocking state would be minimized.

Keywords— Wireless sensor networks, buffer, queuing theory, QoS, multimedia data.

I. INTRODUCTION

Rapid development of Micro Electro-Mechanical Systems (MEMS) led to extensive usage of wireless sensor networks (WSNs) in range of application areas. WSNs consist of wirelessly connected and densely deployed sensor nodes that are small in size and scattered randomly throughout the observed phenomenon area. Sensor nodes are tiny, low-cost, resource limited, energy constrained devices that have sensor boards, processors and transceivers on them. With the sensors embedded to the nodes sense environmental phenomenon, gather observed data and send them to neighboring nodes. Beside it, each node plays a role of a router, where it transfers arrival packets to another neighboring node until they reach to the destination source called a sink [1]. A sink connects a sensor networks with external networks such as internet. Unlike a sensor node, a sink is powerful in terms of continuous

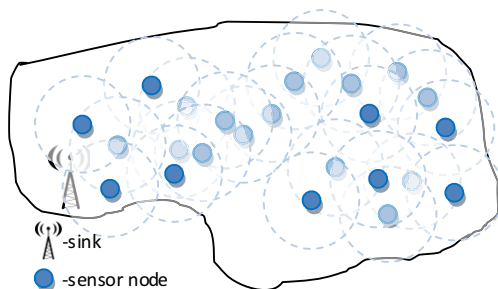


Fig. 1. Wireless sensor networks (WSN) system view

power supply and high computational speed. Data transfer through the nodes is based on multi-hop scenario (Fig. 1).

WSNs have a range of application areas such as: *Machine control service*: In production process, states of all kind of equipments, even small equipments, are very important in order to avoid any machinery accident. Just by equipping small sensor node on the machines, any damages or accidents are detected automatically and are wirelessly transferred to the administrator; *Military surveillances*: WSNs plays important role in military application. Since sensor node is tiny and unnoticed, they are widely used for military purpose, mainly in border control, detection of intruder and armored machines in the battle field; *Patient health control*: To control health of patients in the hospital, such as regular control of blood pressure and etc; *Vehicle control*: In order to determine whether there is a free space in a parking lot, implementation of cheaper sensor node is the best solution; *Noise level control*: In some environmental research practices, such as level of volcanic reservoirs, earthquake signal detection, noise level control is essential factor. Implementation of these sensor nodes in the environmental phenomenon facilities work flow, where nodes detect noise level periodically and transmit them to the sink throughout the nodes in the network.

State-of-the-art sensor nodes have the ability to analyze and filter sensed data and get necessary data in order to send them to the sink, which in turn it reduces a number of retransmissions that result in less energy consumption. It is very important factor to minimize energy usage in designing of both sensor networks and nodes.

Improvement of CMOS (Complimentary Metal Oxide Semiconductor) cameras and microphones extended a range of applications of WSNs, where different traffics classes such as videos, sounds, pictures and real-time streaming were able to transfer via the sensor nodes [2]. These devices were embedded to the sensor nodes in order to capture or to monitor an environment. Unlike traditional WSNs, these traffic classes in multimedia WSNs are required high bandwidth and reliable transfer through the network because of large amount of data size. Although size of video and picture frames taken by CMOS cameras are less than normal wired cameras, they are still quite big for the sensor network because of hardware and energy limitation. In this regard, it is more relevant to reduce quality of picture or video frames in order to decrease file size. However, some applications of WSNs are required high quality. In this scenario, main purpose in designing of the

sensor networks and nodes becomes how to conserve energy and prolong network life.

Sensors are always awake to sense phenomenon and as they gather data, they convey them to the destination node on the bases of following types of request- *a) event-driven*: when the observed environmental experiment happens, sensor starts gathering data and sends them the sink; *b) periodically*: in this type of request, sensors are activated periodically on specific times given in advance; *c) query-based*: sensors act as they get query from an administrator. In this mechanism query is broadcasted by the sink and as the target node gets query, it starts sensing an environmental phenomenon [3].

Unlike mobile ad-hoc wireless network and wired networks such as internet, WSNs required totally different quality of services (QoS) models. In WSNs all QoS measurements designed for routing protocols, path selection, encryption/description and etc should be modeled so that network life would be prolonged. In other words, QoS model in WSNs should involve in it energy consumption. However, this factor is not mandatory one for other type of network, since they have continuous power supply. Furthermore, QoS metrics

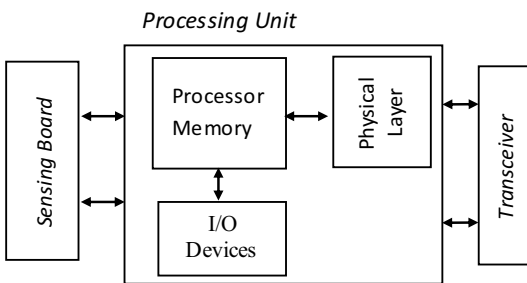


Fig. 2. Basic hardware architecture of a sensor node

varies in WSNs depending on its application area. Basically, additional tasks are required an extra computation, where in turn more energy are consumed. Therefore, QoS models designed for WSNs should have application-specific characteristics [4]. QoS metrics in WSNs can be determined as an effect of different factors, such as path selection, routing protocol, MAC layer protocol, queue management, error correction and etc. Specifically, in video sensor networks quality of images is very important and priority factor in QoS modeling. In general there are several common QoS measurement factors to be considered in WSNs [5]. Main factors of them are *a) energy efficiency* and *b) coverage*:

Energy efficiency: As we mentioned above energy consumption as an essential factor should be involved in most QoS models in WSNs. Sensor network topology are mainly dependent on the battery life of sensor nodes. Network topology is always changed, as any of the node life comes to the end. Therefore, a question of how well manage the power source in order to prolong network life is one of the dominant research objectives in WSN [6].

Coverage: Since sensor nodes are wirelessly connected each other, signal strength plays important role in communication. Because of hardware and battery limitation, sensor nodes cannot send data for long distance; therefore distribution of nodes within the coverage of each node should be proportionally calculated [7].

In order to obtain reliable transfer of quality video or picture files, increasing of hardware specifications and capabilities might be one solution. In this case, to prolong network life additional power supply might be used. However, it will increase the total cost of sensor nodes, therefore overall implementation cost. Furthermore, because of big size of the packet transferring via the sensor networks, data loss becomes much bigger. Sometimes loss of video and voice packet might not seriously affect data recollection in the destination host, since missing information may be asked again. But in case of picture files, whole packets must be delivered to the destination host in order to get original data back. To make trade-off between resource and energy consumption, we proposed queue management system in MAC layer. Within the proposed framework a measurement of probability of blocking of the arrival data is suggested by using queuing theory.

II. SYSTEM BACKGROUND OF MULTIMEDIA WSNs

A wireless sensor nodes consist of power supply, sensor boards, antennas (with embedded transmitters and receivers) microcontrollers, CPU (processor unit), random access memory (RAM) and flash memory for low sized data [8]. In multimedia WSNs, in addition, sensor nodes are embedded with low-cost video cameras or sound recorder. Unlike conventional sensor nodes, multimedia sensor nodes gather video, sound or pictures files instead of sensing data such as temperature, humidity and etc. Video cameras or sound recorder are activated when sensors sense intruders.

Microcontrollers are a “tiny” version of a personal computer. They also consist of memory, I/O devices, processor core on a sing integrated circuit (Fig. 2). Technically, microcontrollers plays essential role in designing sensor nodes. Therefore selection of relevant integrated circuits well suited with other peripheral is essention factor in making trade-off between energy consumption and task computations.

A. Network Model and Buffering Management System

Although there are different standards for WSNs, such as WirelessHART, ISA100, IEEE 1451, ZigBee/802.15.4, common accepted protocol stack is determined in five levels (Fig. 3) [1]:

Application Layer: This layer is the one that runs applications in order to do certain specified tasks;

Transport layer: Decomposition of gathering data into small pieces called segments is taken placed in this layer. Segment are small in size, therefore it is quite easy to send through the network. Destination node reassembles arrival segment to original data back;

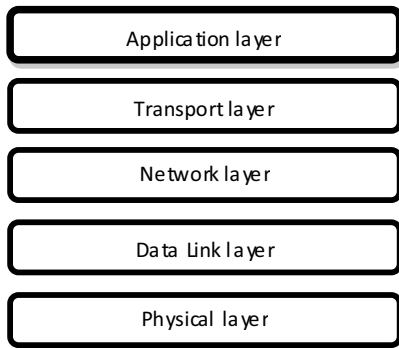


Fig. 3. The WSNs protocol stack

Network layer: Path selection, routing protocol is generated in this layer. In addition, addressing, encapsulation and decapsulation is handled in this layer;

Data Link layer: This layer is divided into two sub layers *Logical Link Control (LLC) layer* and *Medium Access Control (MAC) layer*. LLC layer controls data flow, frame synchronization between the Network layer and MAC layer. MAC layer are responsible for creating medium access between nodes. Common functionalities of MAC layer can be listed as follow: framing, medium, reliability and error control [9]. In this regards, in order to provide better and reliable data transfer between sensor nodes QoS measurement should also be taken consideration on the bases of buffering, queuing, scheduling and transmission of data frames. As proposed in [10] MAC layer QoS can be classified mainly into channel access policies, scheduling and buffer management and error control. In this paper we addressed buffering problem in MAC layer in order to improve QoS in data transmission.

Physical layer: In this layer delivering frames from MAC layer

transforms into wave generated within the transceiver. Following Physical layer elements are necessary to transfer the frames: a) the physical media and connectors; b) encoding and decoding mechanism; c) transmitters and receiver circuitry embedded on sensor nodes.

In this paper we took Physical layer and MAC sub layer together as one “service point” and remaining layers together as second “service point” to determine QoS metric based on the data queue stored in MAC layer.

B. Queuing and Buffering Management System in the MAC layer

One of the best indicators of the packet switching is buffering mechanism in order to store the arrival frames to minimize the data losses. In multimedia WSN it is also very important factor in order to reduce a probability of losses in the sensor nodes during the data transmission. Big size of the data in the multimedia WSN increases loss of data probability. Therefore, scheduling discipline and queuing/buffering management techniques should be well organized in order to manage with the data packet coming from the physical layer to the upper layer. Selection of one datagram among the queued datagrams might be done on the bases of certain mechanism, such as First Come First Serve (FCFS). There are many scheduling discipline in queuing and buffer managements systems. The important issue in queuing management is to provide the best QoS metrics.

To show an importance of queuing model in multimedia WSNs, we illustrated the big picture of the queuing process in Fig. 4. As shown in the figure, Processor Memory is divided into logical part in order to picture the work flow distribution from the perspective of the protocol stack.

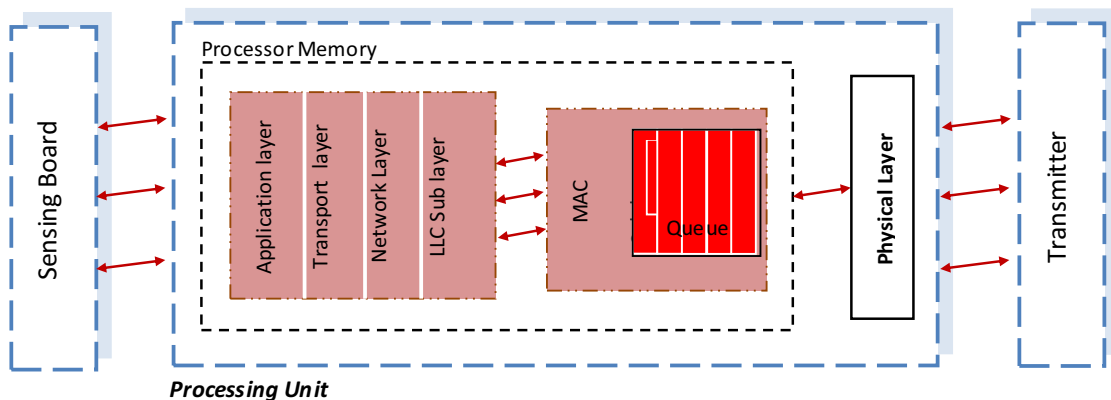


Fig. 4. Queuing process in the Processor Memory

III. RESEARCH FRAMEWORK RESEARCH FRAMEWORK

As we mentioned above, because of resource constraints the sensor nodes cannot serve all the arrival packets at a rapid rate, therefore it result in buffering at the receiver part, namely in the MAC layer. Unlike wired and wireless network, such as internet, ad-hoc, mobile networks, multimedia WSNs requires additional queuing process and scheduling mechanism, MAC protocol because of the low computation speed and energy limitations [11]. Queuing process directly related to the energy consumption. So that data packet loss because of buffer overload leads to retransmission of the same packets several times, which in turn consume additional energy. In this paper we mainly focus on analyzing of queue at MAC layer in order to determine some QoS Metrics.

In the proposed model, we focused on buffering mechanism in the MAC layer and determined QoS Metrics in order to prevent the loss of arrival packets, therefore, to minimize retransmission which in turn results in less power consumption. We analyzed the processor memory in two parts as separate servers. First part of the system is considered as a collection of top four network layers (Application layer, Transport layer, Network layer and LLC layer) with service intensity μ_1 and second part is taken as service point for the MAC layer with service intensity μ_2 (Fig. 5). Arrival packets enter the MAC sublayer through the physical medium (physical layer) and waits in the queue if the system is busy. For simplicity and as a first case, we took bufferless system and showed the QoS metric on arrival data intensity. Based on our result it is possible to determine rate of transmission/receiving and the specification of microcontrollers. In our future works we will consider the buffer and traffic types in our model.

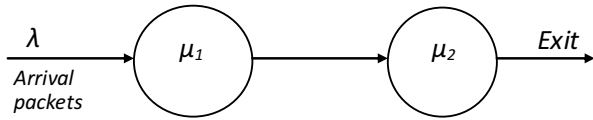


Fig 5. Datagram of data flow in the Queuing Process

In fig. 6 is illustrated the state diagram based on the state spaces given as follow: (0,0), (1,0), (0,1), (1,1), b (block). The state b (blocking) happens when the first server is finished its task, but the second server is still in working regime. If the first serve is not idle, then arrival packet will be dropped.

State probabilities for given state diagram are given in the form *balance equation* as below:

$$\pi(0,0)\lambda = \pi(0,1)\mu_2 \quad (1)$$

$$\pi(1,0)\mu_1 = \pi(0,0)\lambda + \pi(1,1)\mu_2 \quad (2)$$

$$\pi(0,1)(\lambda + \mu_2) = \pi(1,0)\mu_1 + \pi(b)\mu_2 \quad (3)$$

$$\pi(1,1)(\mu_1 + \mu_2) = \pi(0,1)\lambda \quad (4)$$

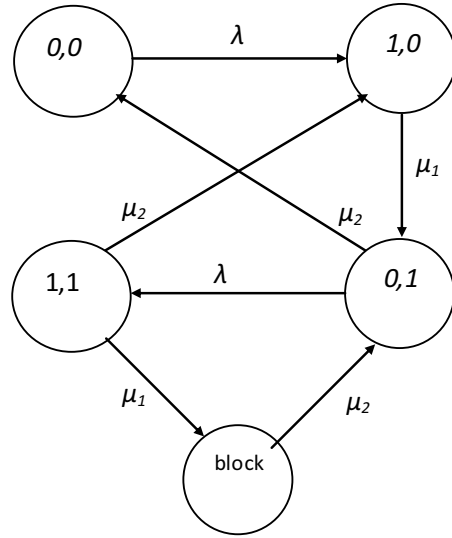


Fig. 6. The state diagram of system

$$\pi(b)\mu_2 = \pi(1,1)\mu_1 \quad (5)$$

Normalizing condition for the given model is:

$$\pi(0,0) + \pi(0,1) + \pi(1,0) + \pi(1,1) + \pi(b) = 1 \quad (6)$$

For analyzing the model we took two cases a) general case, where $\mu_1 \neq \mu_2$ and b) special case, where $\mu_1 = \mu_2$.

In general case, computation speed of the CPU should be faster than receiving and service speed at the MAC layer. Other vice, packets flowing from the MAC layer to the up layers either will be loss or should wait in the queue. Since in our model we consider bufferless system in each part, we can propose that $\mu_2 > \mu_1$. Nevertheless our model is designed to all cases under the condition of $\mu_1 \neq \mu_2$. Using mathematical calculation we can formulize the state probabilities π as follow:

$$\pi(0,1) = \frac{\lambda}{\mu_2} \pi(0,0) = \rho_2 \pi(0,0), \quad \rho_2 = \frac{\lambda}{\mu_2} \quad (7)$$

$$\pi(1,1) = \frac{\lambda}{\mu_1 + \mu_2} \pi(0,1) = \frac{\lambda}{\mu_1 + \mu_2} \rho_2 \pi(0,0) = \rho \rho_2 \pi(0,0),$$

where $\rho = \frac{\lambda}{\mu_1 + \mu_2}$ (8)

$$\pi(b) = \frac{\mu_1}{\mu_2} \pi(1,1) = \frac{\mu_1}{\mu_2} \rho \rho_2 \pi(0,0) \quad (9)$$

$$\pi(1,0) = \frac{\lambda}{\mu_1} \pi(0,0) + \frac{\mu_2}{\mu_1} \pi(1,1) =$$

$$\rho_1 \pi(0,0) + \frac{\mu_2}{\mu_1} \rho \rho_2 \pi(0,0) = \left(\rho_1 + \frac{\mu_2}{\mu_1} \rho \rho_2 \right) \pi(0,0) =$$

$$= \left(\rho_1 + \frac{\lambda}{\mu_1} \rho \right) \pi(0,0) = (\rho_1 + \rho_1 \rho) \pi(0,0) = \rho_1 (1 + \rho) \pi(0,0),$$

where $\rho_1 = \frac{\lambda}{\mu_1}$ (10)

$$\pi(0,0) = \left(1 + \rho_2 + \rho \rho_2 + \frac{\mu_1}{\mu_2} \rho \rho_2 + \rho_1 + \rho \rho_1\right)^{-1} = \left(1 + (1 + \rho)(\rho_2 + \rho_1) + \frac{\mu_1}{\mu_2} \rho \rho_2\right)^{-1} \quad (11)$$

As a QoS metrics, probability of blocking PB can be found as below:

$$PB = \pi(1,0) + \pi(1,1) + \pi(b) \quad (12)$$

In special case, we analyzed the case where $\mu_1 = \mu_2$, $\rho = \frac{\lambda}{\mu}$. In this case state probabilities π can be given as follow:

$$\pi(0,1) = \frac{\lambda}{\mu} \pi(0,0) = \rho \pi(0,0) \quad (13)$$

$$\pi(1,1) = \frac{\lambda}{2\mu} \pi(0,1) = \frac{\rho}{2} \pi(0,1) = \frac{\rho^2}{2} \pi(0,0) \quad (14)$$

$$\pi(b) = \pi(1,1) = \frac{\rho^2}{2} \pi(0,0) \quad (15)$$

$$\pi(1,0) = \frac{\lambda}{\mu} \pi(0,0) + \pi(1,1) = \rho \pi(0,0) + \pi(0,0) = \left(\rho + \frac{\rho^2}{2}\right) \pi(0,0) \quad (16)$$

$$\begin{aligned} \pi(0,0) &= \left(1 + \rho + \frac{\rho^2}{2} + \frac{\rho^2}{2} + \frac{\rho^2}{2} + \rho\right)^{-1} = \\ &= \left(1 + 2\rho + \frac{3\rho^2}{2}\right)^{-1} \end{aligned} \quad (17)$$

QoS metrics, probability of blocking for this case is:

$$\begin{aligned} PB &= \pi(1,0) + \pi(1,1) + \pi(b) = \left(\frac{\rho^2}{2} + \rho + \frac{\rho^2}{2} + \right. \\ &\left. \frac{\rho^2}{2}\right) \pi(0,0) = \frac{\frac{3\rho^2}{2} + \rho}{1 + 2\rho + \frac{3\rho^2}{2}} = \frac{3\rho^2 + 2\rho}{2 + 4\rho + 6\rho^2} \end{aligned} \quad (18)$$

Based on the exact formulas of state probabilities, we did numerical analyses for both cases.

IV. CONCLUSION

The simplest case of queuing/buffering management system - bufferless case gives us how intensive rates should be chosen in order to reduce data loss based on the result of the probability of blocking. The numerical analyze of the given model will published in our future work. Based on our result it is possible to determine desirable rate of transmission/receiving and the specification of microcontrollers. We estimated optimal values for each server and proposed desirable service intensity values for sensor node in order to minimize data loss. Probability of blocking gives us clear picture how system specification should be chosen so that blocking state would be minimized.

- [1] I.Akyildiz, W. Su, Y. Sankarasubramaniam, E. Cayirci, "Wireless sensor networks: A survey", *Computer Networks*, 38 (4) (2002) 393–422.
- [2] I.F. Akyildiz, T. Melodia and K. R. Chowdhury, "A survey on wireless multimedia sensor networks", *Computer Network*, vol. 51, pp. 921–960, 2007.
- [3] H. C. Lee, H. Guyennet and N. Zerhouni, "Redundant Communication Avoidance for Event-Driven Wireless Sensor Network", *IJCSNS International Journal of Computer Science and Network Security*, VOL.7 No.3, pp. 193-200, March 2007
- [4] Z.A. Boukerche; R.B. Araujo; Villas, L. A, "Wireless Actor and Sensor Networks QoS-Aware Routing Protocol for the Emergency Preparedness Class of Applications", *In Proc. 31st IEEE Confon Local Computer Networks*, Tampa, FL, 2006; pp. 832-839
- [5] E. Felemba; C. Lee; E. Ekici, "MMSPEED: Multipath Multi-SPEED protocol for QoS guarantee of reliability and timeliness in wireless sensor networks", *IEEE Transactions on Mobile Computing*, 2007, 5(6), 738-754
- [6] N. Heo and P. K. Varshney, "Energy-Efficient Deployment of Intelligent Mobile Sensor Networks", *IEEE Transactions on Systems, Man and Cybernetics*, Vol. 35, No. 1, Jan 2005.
- [7] L. Hu and D. Evans, "Localization for Mobile Sensor Networks", *Proceedings of the 10th Annual International Conference on Mobile Computing and Networking (ACM Mobicom '04)*, Philadelphia, PA, USA, 26 Sep – 1 Oct, 2004.
- [8] J. Hill, R. Szewczyk, A. Woo, S. Hollar, D. Culler, and K. Pister, "System Architecture Directions for Networked Sensors", *ASPLOS*, November 2000.
- [9] K. Kredb, P. Mohapatra, "Medium Access Control in WSNs", *Computer Networks*, Vol 2, June, 2006.
- [10] S. A. Khan, S. A. Arshad, "QoS Provisioning Using Hybrid FSO-RF Based Hierarchical Model for Wireless Multimedia Sensor Networks", *(IJCSIS) International Journal of Computer Science and Information Security*, Vol. 4, No. 1 & 2, 2009.
- [11] E. Magli, M. Mancin, and L. Merello, "Low-Complexity Video Compression for Wireless Sensor Networks", *Proc. ICME 2003*, vol. 3, pp. 585-588, July 2003.

Coded OFDM Wireless Systems with Generalized Prefix

Hakan Doğan, Hakan Yıldız, *Graduate Member, IEEE*, Todor Cooklev, *Senior Member, IEEE*, Yusuf Acar

Abstract—In the case of orthogonal frequency-division multiplexing (OFDM) systems, associated subcarriers can experience deep fades. When a cyclic prefix (CP) is used, CP-OFDM systems suffer from deep fading effects. With the channel coding and interleaving, coded OFDM provides a robust communication link for wireless channels. Coded CP-OFDM (COFDM) is one of the widely used transmission techniques for overcoming the deep fading of the channel and it tries to minimize deep fading problem by the use of forward error correction (FEC) techniques. Since it is known that the Coded CP-OFDM is not an exact solution to overcome the deep fading effect that degrades significantly the system performance, Coded generalized prefix (GP)-OFDM is proposed and also compared with coded CP-OFDM.

Index Terms—Generalized Prefix, OFDM, ISI, Fading, FEC, Viterbi, Convolution Code, Cyclix Prefix.

I. INTRODUCTION

Bandlimited wireless communication channels exhibit significant variations in gain and phase along different frequencies. This fact truncates the capacity of the channel. In the case of orthogonal frequency-division multiplexing (OFDM) systems, associated subcarriers can experience deep fades. In the extreme case these deep fades may turn into spectral nulls. When a cyclic prefix (CP) is used, these spectral nulls are known to significantly limit the bit error rate performance.

CP-OFDM systems suffer from deep fading effects. In literature, there are several approaches which are called Zero Padding (ZP)-OFDM [1], [2], pseudo random postfix (PRP)-OFDM [3], known symbol padding (KSP) OFDM [4], [5], unique word (UW) OFDM [6] that are proposed to deal with the deep fading problem. Practical OFDM systems [7] simply accept this difficulty, and try to minimize it by the use of forward error correction (FEC) techniques [8], which is called Coded CP-OFDM.

Recently, a new prefix like CP but weighted by a complex number called generalized prefix (GP) OFDM is proposed and it is shown that it has the lowest BER among the other prefix construction techniques, considering all other parameters identical [9]. In GP-OFDM, the prefix that is used is

Copyright (c) 2012 IEEE. Personal use of this material is permitted. However, permission to use this material for any other purposes must be obtained from the IEEE by sending a request to pubs-permissions@ieee.org.

Hakan Doğan is with the Department of Electrical and Electronics Engineering, Istanbul University, 34320, Avcılar, Istanbul, Turkey (e-mail: hdogan@istanbul.edu.tr).

Todor Cooklev is with the Wireless Technology Center, Indiana University-Purdue University Fort Wayne (IPFW), USA (e-mail: cooklevt@ipfw.edu).

Hakan Yıldız is with Ericsson, 34398, Maslak, Istanbul, Turkey (e-mail: hakan.yildiz@ericsson.com).

Yusuf is with Department of Electronics Engineering, Istanbul Kultur University Bakirkoy, 34156, Istanbul, Turkey (e-mail: y.acar@iku.edu.tr)

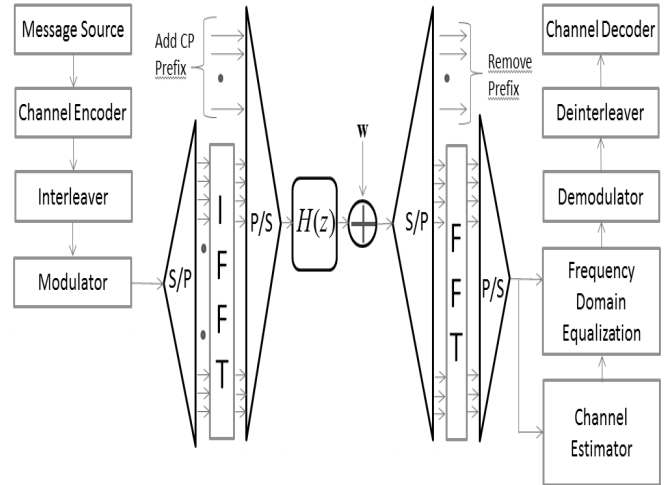


Fig. 1. Conventional Cyclic Prefix (CP) Based Coded-OFDM System

not random; it is determined using a computationally simple optimization routines. This prefix shifts the phases of the multipath components and effectively changes the wireless channel experienced by the OFDM system into a different channel. In particular, a wireless channel that has deep fades or spectral nulls is transformed into a channel with fades that are less deep or no spectral nulls. Furthermore, it is also demonstrated that GP-OFDM is also superior in the presence of channel estimation errors versus CP-OFDM while it is overall computationally simple and allows the use of cross-correlation or autocorrelation based synchronization methods as CP-OFDM.

Coded CP-OFDM (COFDM) is one of the widely used transmission techniques for overcoming the deep fading of the channel. The basic idea of coded OFDM is to encode input data and interleave the coded symbols. The interleaved symbols are split into several subchannels to achieve frequency diversity. Even though the uncoded symbol error rate is high for the subcarriers with low channel gains, with the channel coding and interleaving it is possible to correct the errors in the low gain channels. With the channel coding and interleaving, coded OFDM provides a robust communication link for wireless channels. However, it is proved that the Coded CP-OFDM is not an exact solution to overcome the deep fading effect that degrades significantly the system performance. Therefore, in this paper, GP-OFDM is combined with the channel coding and compared to the Coded-CP-OFDM systems since Coded CP OFDM and Coded GP OFDM have not been compared

yet. Both schemes are compared to prove whether GP OFDM is superior to CP OFDM also in channel coding environment.

II. CODED-CP-OFDM

Let us consider a convolutionally coded OFDM system with N subcarriers and available bandwidth $B = 1/T_s$ where T_s is the sampling period. A given sampling period is divided into N subchannels by equal frequency spacing $\Delta f = B/N$. At the transmitter, a block of the information bit sequence $\{b(u), u = 0, 1, \dots\}$ is applied to a convolutional encoder of rate R . The binary encoder output is fed into a block interleaver and then directly mapped into the data symbol vector c_k according to the modulation format employed. In this case, for one OFDM symbol, the N frequency domain constellation symbol c_k is given as follows;

$$\mathbf{c} = [c_1, c_2, \dots, c_N]^T \quad (1)$$

The time domain equivalent is obtained by the N point inverse DTF as follows

$$\mathbf{s} = \mathbf{F}^* \mathbf{c} = [s_1, s_2, \dots, s_N]^T \quad (2)$$

Then, the cyclic prefix is inserted as $\mathbf{S} = \mathbf{G} \cdot \mathbf{s}$, where \mathbf{G} has $N + P$ rows and N columns:

$$\mathbf{G} = \begin{bmatrix} \frac{O_{P \times N-P} \mid I_P}{I_N} \end{bmatrix} \quad (3)$$

it yields the transmitted signal as

$$\mathbf{s}' = \begin{bmatrix} \underbrace{S_N \ S_{N-1} \ \dots \ S_{N-P+1}}_P \mid \underbrace{S_1, S_2, \dots, S_N}_N \end{bmatrix}^T \quad (4)$$

The received signal over wireless channel

$$\mathbf{r}' = \mathbf{H} \mathbf{s}' + \mathbf{n} \quad (5)$$

where

$$\mathbf{H} = \begin{bmatrix} \mathbf{h}_1 & & & & & \\ \mathbf{h}_2 & \mathbf{h}_1 & & & & \\ \vdots & \mathbf{h}_2 & & & & \\ \mathbf{h}_L & \vdots & \ddots & & \mathbf{h}_1 & \\ 0 & \mathbf{h}_L & & \ddots & \mathbf{h}_2 & \\ \vdots & \vdots & & & \mathbf{h}_{L-1} & \\ 0 & 0 & & & \mathbf{h}_L & \end{bmatrix}_{(N+P+L-1) \times (P+N)} \quad (6)$$

The wideband wireless channel is modeled with respect to the baseband by a FIR model with L channel coefficients.

Cyclic-prefix is removed by discarding the P values of the received signal and DTF is applied.

$$\mathbf{r} = \mathbf{F} \mathbf{M} \mathbf{r}' = \mathbf{F} \mathbf{M} \mathbf{H} \mathbf{G} \mathbf{F}^* \mathbf{c} + \mathbf{w} \quad (7)$$

where

$$\mathbf{M} = \begin{bmatrix} O_{N \times P} & I_N & O_{N \times L-1} \end{bmatrix} \quad (8)$$

Simplified form can be written as

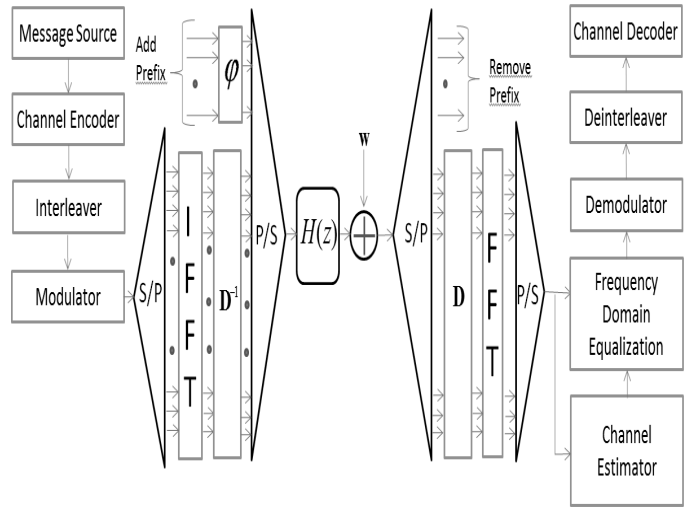


Fig. 2. Proposed Generalized Prefix (GP) Based Coded-OFDM System

$$\mathbf{r} = \bar{\mathbf{H}} \mathbf{c} + \mathbf{w} \quad (9)$$

where $\bar{\mathbf{H}} = \mathbf{F} \mathbf{M} \mathbf{H} \mathbf{G} \mathbf{F}^*$ and it is a diagonal matrix because it is assumed that the channel is static for one OFDM symbol duration. Channel frequency response for CP-OFDM system could written as follows;

$$\bar{\mathbf{H}} = \text{diag}[H(0), H(1), \dots, H(N-1)] \quad (10)$$

where

$$\mathbf{H}(k) = \mathbf{H}(e^{jwk}) = \sum_{n=0}^{N-1} \mathbf{h}[n] e^{-jwkn} \quad (11)$$

III. CODED-GP-OFDM

After IDFT process, two steps are done to add generalized prefix;

Step 1: Apply \mathbf{D}^{-1} matrix

$$\mathbf{S} = \mathbf{D}^{-1} \mathbf{s} \quad (12)$$

where \mathbf{D} matrix

$$\mathbf{D} = \begin{bmatrix} 1 & & & & & \\ & \psi & & & & \\ & & \psi^2 & & & \\ & & & \ddots & & \\ & & & & \psi^{N-1} & \end{bmatrix} \quad (13)$$

Step 2: Add prefix prefix

$$\mathbf{s}' = \mathbf{G}_\psi \mathbf{S} \quad (14)$$

where

$$\mathbf{G}_\psi = \begin{bmatrix} O_{P \times (N-P)} & \psi^N \cdot I_P \\ & I_N \end{bmatrix} \quad (15)$$

in this case we have transmitted signal as follows

$$\mathbf{s}' = \mathbf{G}_\psi \mathbf{D}^{-1} \mathbf{s} \quad (16)$$

Similarly, prefix is removed by \mathbf{M} matrix at the receiver. After application of \mathbf{D}^{-1} and \mathbf{F} matrices we have

$$\mathbf{r} = \mathbf{FDMHG}_\psi \mathbf{D}^{-1} \mathbf{F}^* \mathbf{c} + \mathbf{w} \quad (17)$$

Equation can be simplified as (18)

$$\mathbf{r} = \bar{\mathbf{H}}_\psi \mathbf{c} + \mathbf{w} \quad (18)$$

where $\bar{\mathbf{H}}_\psi = \mathbf{FDMHD}^{-1} \mathbf{G}_\psi \mathbf{F}^*$. It is shown in [9], the channel frequency response could be written for GP-OFDM system as follows

$$\bar{\mathbf{H}}_\psi = \text{diag}[H_\psi(0), H_\psi(1), \dots, H_\psi(N-1)] \quad (19)$$

where

$$\mathbf{H}_\psi(k) = \sum_{n=0}^{N-1} \psi^n \mathbf{h}[n] e^{-j\omega kn} \quad (20)$$

It was shown that better frequency responses can be achieved by the optimization of ψ [9].

IV. SELECTION OF ψ FOR GP-OFDM SYSTEM

The transfer function and frequency response of the wireless channel could be written as follows for CP-OFDM system;

$$H(z) = h[0] + h[1]z^{-1} + \dots + h[L-1]z^{-(L-1)}. \quad (21)$$

$$H(e^{j\omega}) = \sum_{n=0}^{N-1} h[n] e^{-j\omega n} \quad (22)$$

It is shown that a defined shift in frequency domain of the channel response could be done for GP-OFDM systems [9]. In this case, the wireless channel is transformed into following equivalent transfer function and the frequency response

$$H_\psi(z) = h[0] + \psi h[1]z^{-1} + \psi^2 h[2]z^{-2} + \dots + \psi^{L-1} h[L-1]z^{-(L-1)} \quad (23)$$

$$H_\psi(e^{j\omega}) = \sum_{n=0}^{N-1} \psi^n h[n] e^{-j\omega n} = H(e^{j(\omega-\alpha)}), \quad (24)$$

As a conclusion, the wireless channel may have less severe deep fading that supplies better BER by the use of different $\psi = e^{j\alpha}$ values where $\alpha \in [0, \frac{2\pi}{N}]$.

A. Optimization

Searching for the optimal value of α (therefore, ψ) could be done by the usage of Max-min approach that makes the minimum values of $H_\psi[k]$ (deep fades) to be as large as possible as follows [10];

$$\alpha^* = \max_{\alpha \in [0, \frac{2\pi}{N}]} O_c(\psi)|_{\psi=e^{j\alpha}}. \quad (25)$$

where the objective cost function is defined as $O_c(\psi) = \min_k |H_\psi[k]|$

The proposed technique requires ψ to be determined by using channel state information (CSI). Although ψ can be computed by evaluating CSI information at the transmitter,

this technique requires more bandwidth. Therefore by using pilot-tones, which are embedded in the transmitted downlink baseband signal, ψ can be calculated at the receiver and only the value of ψ is sent to the transmitter. In this case, feedback can be provided with minimum bandwidth and more compatible to the channel environment.

B. Golden Section

The maximum of a unimodal function given in Eq.(25) could be find by successively narrowing the range of values inside which the extremum is known to exist. The one-dimensional optimization problem needs searching a bracketed interval can be solved by means of numerical techniques such as the golden section search method without resorting to derivatives [11].

In other words, the golden section search is a very efficient algorithm for finding out the extreme (minimum or maximum) of an objective function with unimodal when the calculation of derivative is unable or it needs more complexity.

Let us assume that $[a, b]$ is a given interval where the sought minimum is located. In this case, the points p and q can be calculated as follows:

$$\begin{aligned} p &= b - (b-a)\Phi, \\ q &= a + (b-a)\Phi, \end{aligned} \quad (26)$$

where $\Phi = \frac{\sqrt{5}-1}{2}$ is the golden ratio conjugate. After evaluating the function $O_c(\cdot)$ at points p and q , a new search interval is established according to whether $O_c(e^{jp}) \leq O_c(e^{jq})$ or not. Details of the algorithm will not be presented here since the golden Section algorithm has been well studied in literature [12].

V. EQUALIZATION

There are many types of equalizers such as linear equalizer, decision feedback equalizer (DFE), maximum a posteriori probability (MAP) equalizer, soft-output Viterbi (MLSE) equalizer in the literature [13], [14]. In this section, a simple frequency domain linear equalizer called the zero forcing (ZF) is employed.

Zero forcing equalizer applies the inverse of the frequency response of the channel to remove the effect of channel from the received signal as follows;

$$\mathbf{z} = \hat{\mathbf{H}}_\psi^{-1} \mathbf{r} \quad (27)$$

where $\hat{\mathbf{H}}_\psi^{-1}$ is the estimated channel frequency response.

VI. DECODING

The Viterbi Decoder that enables the maximum likelihood (an asymptotically optimal) decoding to recover the convolutionally encoded data [15]. It also has the advantage of a fixed decoding time and is well suited to hardware implementations. Therefore, it is employed in many Forward Error Correction (FEC) applications and standards, such as LTE, IEEE 802.11, IEEE 802.16, Hiperlan [16].

In this paper, the Viterbi decoder is used to decode the equalized signal in Eq.(27). The decoder needs the equalized

signal and parameters that are used at the transmitter such as code rate, constraint length, the generator polynomials. Hard or soft decoding may be done in the Viterbi decoder.

VII. SIMULATION RESULTS

In this section, we perform computer simulations to compare the BER performance of Coded GP-OFDM and Coded CP-OFDM systems. The system operates with a bandwidth of 5 MHz and is divided into 512 tones ($N = 512$) with a total symbol period of $108.8 \mu\text{s}$, of which $6.4 \mu\text{s}$ is allocated to the prefix (32 samples). The wireless channels between the mobile antenna and the receiver antenna are modeled based on a realistic channel model determined by the COST-207 project in which Typical Urban (TU) channel model is considered [17]. The information symbols are QPSK modulated and the coding scheme chosen for simulations is the (138, 158) convolutional code with rate 1/2. Since burst errors deteriorate the performance of the coding scheme, the output sequence from the encoder is interleaved with a 32×32 block interleaver to spread the consequences of a local notch in the transfer function over the code sequence.

At the receiver, it is assumed that channel frequency response is estimated with all pilot based channel estimation in the training mode and then equalized by the zero forcing equalizer. Later, the equalized signal is demodulated and hard Viterbi decoding algorithm (the decoder expects binary input values) is employed to decode binary data.

Fig. 1 compares the BER performance of the proposed coded GP-OFDM and coded CP-OFDM systems as a function of energy per bit to noise power ratio (E_b/N_0) where N_0 is equal to σ_w^2 . It is observed that coded-GP-OFDM outperforms coded CP-OFDM and using the coded GP-OFDM system leads to a substantial performance advantage. In particular, it outperforms CP-OFDM by about 2.8 dB at $\text{BER} = 10^{-4}$ in Fig.1 for the 12-tap TU channel.

VIII. CONCLUSIONS

In this paper Coded GP-OFDM is proposed to overcome the spectral nulls which possibly exist in fading channels in vehicular communications environment and compared with Coded CP-OFDM. It is shown that the proposed system outperforms the conventional coded OFDM system over channels with deep fading. When spectral nulls occur, the proposed scheme improves the BER performance of the system. From the simulation results it can be concluded that the performance of the OFDM-wireless communication system over fading channels can be improved when Coded GP-OFDM is applied. Therefore, the proposed technique may be considered for OFDM based wireless indoor standards where channel variations could be tracked easily.

REFERENCES

- [1] A. Scaglione, G. Giannakis, and S. Barbarossa, "Redundant filterbank precoders and equalizers. i. unification and optimal designs," *Signal Processing, IEEE Transactions on*, vol. 47, no. 7, pp. 1988–2006, 1999.
- [2] B. Muquet, Z. Wang, G. B. Giannakis, M. de Courville, and P. Duhamel, "Cyclic prefixing or zero padding for wireless multicarrier transmissions?" *IEEE Transactions on Communications*, vol. 50, no. 12, pp. 2136–2148, 2002.

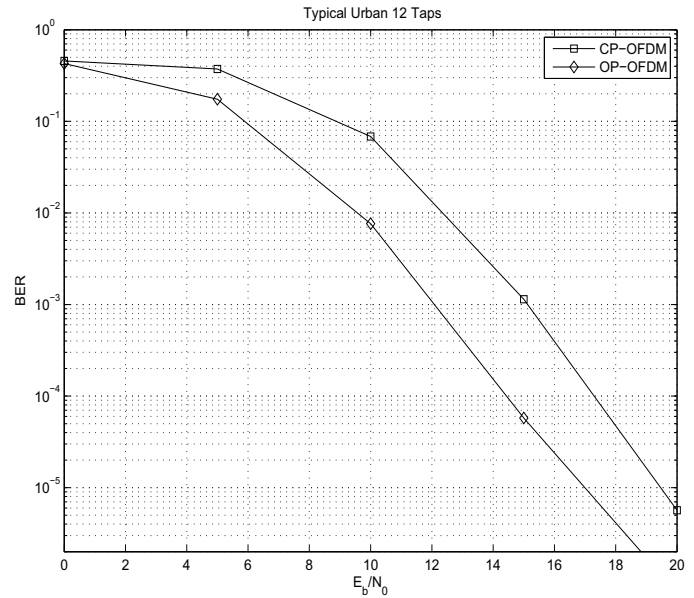


Fig. 3. BER performance of the GP-OFDM and CP-OFDM with FEC decoding for the 12-tap TU channel

- [3] M. Muck, M. de Courville, and P. Duhamel, "A pseudorandom postfix OFDM modulator—Semi-blind channel estimation and equalization," *IEEE Transactions on Signal Processing*, vol. 54, no. 3, pp. 1005–1017, 2006.
- [4] L. Deneire, B. Gyselinckx, and M. Engels, "Training sequence versus cyclic prefix—a new look on single carrier communication," *Communications Letters, IEEE*, vol. 5, no. 7, pp. 292–294, 2001.
- [5] M. Huemer, C. Hofbauer, and J. B. Huber, "Unique word prefix in SC/FDE and OFDM: A comparison," in *IEEE GLOBECOM Workshops*, Miami, USA, Dec. 2010, pp. 1296–1301.
- [6] M. Huemer, C. Hofbauer, and J. Huber, "Complex number RS coded ofdm with systematic noise in the guard interval," in *Signals, Systems and Computers (ASILOMAR), 2010 Conference Record of the Forty Fourth Asilomar Conference on*. IEEE, 2010, pp. 1023–1028.
- [7] I. P802.11n, Part 11: *Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications: Amendment 5: Enhancements for Higher Throughput*. IEEE-802.11 WG, Oct. 2009.
- [8] Z. Wang and G. B. Giannakis, "Linearly precoded or coded OFDM against wireless channel fades?" in *Proceedings of the Third IEEE Workshop Signal Processing Advances for Wireless Communications*, Mar. 2001.
- [9] T. Cooklev, H. Dogan, R. Cintra, and H. Yildiz, "A generalized prefix construction for ofdm systems over quasi-static channels," *Vehicular Technology, IEEE Transactions on*, vol. 60, no. 8, pp. 3684–3693, oct. 2011.
- [10] Z. Drezner, "On minimax optimization problems," *Mathematical Programming*, vol. 22, no. 1, pp. 227–230, 1982.
- [11] S. S. Rao, *Engineering Optimization: Theory and Practice*. Wiley-Interscience, 2009.
- [12] K. Deb, *Optimization for engineering design: Algorithms and examples*. Prentice-Hall of India, 2005.
- [13] J. Proakis and M. Salehi, *Fundamentals of communication systems*. Pearson Education, 2007.
- [14] J. Proakis, M. Salehi, and G. Bauch, *Contemporary communication systems using MATLAB*. Thomson Engineering, 2012.
- [15] A. Viterbi, "Error bounds for convolutional codes and an asymptotically optimum decoding algorithm," *Information Theory, IEEE Transactions on*, vol. 13, no. 2, pp. 260–269, 1967.
- [16] A. Doufexi, S. Armour, M. Butler, A. Nix, D. Bull, J. McGeehan, and P. Karlsson, "A comparison of the hiperlan/2 and ieee 802.11 a wireless lan standards," *Communications Magazine, IEEE*, vol. 40, no. 5, pp. 172–180, 2002.
- [17] J. B. Hanzo and L. Hanzo, *Third-generation Systems and Intelligent Wireless Networking: Smart Antennas and Adaptive Modulation*. Wiley-IEEE Press, 2002.

Evolution Mobile Wireless Communication And LTE Networks

Tinatin Mshvidobadze

Tech. doctor of science- Professor

7/33 Tskhinvali Street -Gori (Georgia) University

tinikomshvidobadze@gmail.com

Abstract

Mobile communications systems revolutionized the way people communicate, joining together communications and mobility. A long way in a remarkably short time has been achieved in the history of wireless. Evolution of wireless access technologies is about to reach its fourth generation (4G). Looking past, wireless access technologies have followed different evolutionary paths aimed at unified target: performance and efficiency in high mobile environment.

The first generation (1G) has fulfilled the basic mobile voice, while the second generation (2G) has introduced capacity and coverage. This is followed by the third generation (3G), which has quest for data at higher speeds to open the gates for truly “mobile broadband” experience, which will be further realized by the fourth generation (4G).

The Fourth generation (4G) will provide access to wide range of telecommunication services, including advanced mobile services, supported by mobile and fixed networks, which are increasingly packet based, along with a support for low to high mobility applications and wide range of data rates, in accordance with service demands in multiuser environment.

This paper provides a high level overview of the evolution of Mobile Wireless Communication Networks from 3G to 4G.

Is described LTE(Long Term Evolution) a fourth generation (4G) mobile network technology.

Keywords: Mobile Wireless Communication Networks, 3G, 4G, Mobile Broadband, Long Term Evolution.

Introduction

The last few years have witnessed a phenomenal growth in the wireless industry, both in terms of mobile technology and its subscribers. There has

been a clear shift from fixed to mobile cellular telephony, especially since the turn of the century.

By the end of 2010, there were over four times more mobile cellular subscriptions than fixed telephone lines. Both the mobile network operators and vendors have felt the importance of efficient networks with equally efficient design. This resulted in Network Planning and optimization related services coming in to sharp focus [1].

With all the technological advances, and the simultaneous existence of the 2G, 2.5G and 3G networks, the impact of services on network efficiency have become even more critical. Many more designing scenarios have developed with not only 2G networks but also with the evolution of 2G to 2.5G or even to 3G networks. Along with this, inter-operability of the networks has to be considered.

1G refers to analog cellular technologies; it became available in the 1980s. 2G denotes initial digital systems, introducing services such as short messaging and lower speed data. CDMA2000 1xRTT and GSM are the primary 2G technologies, although CDMA2000 1xRTT is sometimes called a 3G technology because it meets the 144 kbps mobile throughput requirement. EDGE, however, also meets this requirement. 2G technologies became available in the 1990s. 3G requirements were specified by the ITU as part of the International Mobile Telephone 2000 (IMT-2000) project, for which digital networks had to provide 144 kbps of throughput at mobile speeds, 384 kbps at pedestrian speeds, and 2 Mbps in indoor environments. UMTS-HSPA and CDMA2000 EV-DO are the primary 3G technologies, although recently WiMAX was also designated as an official 3G technology. 3G technologies began to be deployed last decade.

The ITU has recently issued requirements for IMT-Advanced, which constitutes the official definition of 4G. Requirements include operation in up-to-40 MHz radio channels and extremely high spectral efficiency. The ITU recommends operation in upto- 100 MHz radio channels and

peak spectral efficiency of 15 bps/Hz, resulting in a theoretical throughput rate of 1.5 Gbps. Previous to the publication of the requirements, 1 Gbps was frequently cited as a 4G goal. No available technology meets these requirements yet. It will require new technologies such as LTE-Advanced (with work already underway) and IEEE 802.16m. Some have tried to label current versions of WiMAX and LTE as “4G”, but this is only accurate to the extent that such designation refers to the general approach or platform that will be enhanced to meet the 4G requirements. With WiMAX and HSPA significantly outperforming 3G requirements, calling these technologies 3G clearly does not give them full credit, as they are a generation beyond current technologies in capability. But calling them 4G is not correct. Unfortunately, the generational labels do not properly capture the scope of available technologies and have resulted in some amount of market confusion [2].

The long-term evolutionary access technology called LTE (Long Term Evolution) is quickly becoming the network technology of choice for 4G deployments around the world. As user demand for mobile broadband services continues to rise, LTE and its ability to cost-effectively provide very fast, highly responsive mobile data services appears to be the right technology at the right time. By 2014, Juniper Research predicts revenues from LTE mobile broadband subscribers will exceed \$70 billion globally, with 300 million worldwide subscribers by 2015. As of August 2010, according to the Global mobile Suppliers Association (GSA), there are 101 firm LTE network deployments planned or in progress in 41 countries around the world. By the end of 2010, GSA anticipates around 22 LTE networks in commercial service. In addition to these statistics, another 31 operators are currently engaged in various LTE pilot trials and technology tests, which when added to the above translates to 132 operators in 56 countries now investing in LTE. For many operators, LTE represents a significant shift from legacy mobile systems as the first all-Internet Protocol (IP) network technology and will impact the way networks are designed, deployed, and managed. Mobile operators will need to deal with specific challenges associated with LTE, such as interoperability with legacy and other 4G systems, ensuring end-to-end network QoS and high-quality service delivery, and interaction with

IMS for the delivery of multimedia services and voice.

Mobile Network Evolution

LTE is a Fourth Generation (4G) mobile network technology. Mobile networks have evolved through a series of innovations to meet the ever-growing demand for wireless services, beginning with the analog cellular networks introduced almost 30 years ago.

First Generation Cellular Networks

All of the First Generation (1G) mobile systems provided voice services based on analog radio transmission techniques. These first generation technologies used Frequency Division Multiple Access (FDMA) which had inherent limitations in the use of radio channels, and used circuit-switched technologies in the network core.

Second Generation Cellular Network

Second Generation (2G) mobile systems are characterized by digitization and compression of speech. This allowed many more mobile users to be accommodated in the radio spectrum through either time (GSM) or code (IS-95 CDMA) multiplexing.

The Third-generation (WCDMA in UMTS, CDMA2000 & TD-SCDMA)

In EDGE(Enhanced Data rates in GSM Environment), high-volume movement of data was possible, but still the packet transfer on the air-interface behaves like a circuit switch call. Thus part of this packet connection efficiency is lost in the circuit switch environment. Moreover, the standards for developing the networks were different for different parts of the world. Hence, it was decided to have a network which provides services independent of the technology platform and whose network design standards are same globally. Thus, 3G was born. The International Telecommunication Union (ITU) defined the demands for 3G mobile networks with the IMT-2000 standard. An organization called 3rd Generation Partnership Project (3GPP) has continued that work by defining a mobile system that fulfills the IMT-2000 standard. In Europe it was called UMTS (Universal Terrestrial Mobile System), which is ETSI-driven. IMT2000 is the ITU-T name for the third generation system, while cdma2000 is the name of the American 3G

variant. WCDMA is the air-interface technology for the UMTS.

The main components includes BS (Base Station) or node B, RNC (Radio Network Controller), apart from WMSC (Wideband CDMA Mobile Switching Centre) and SGSN/GGSN. 3G networks enable network operators to offer users a wider range of more advanced services while achieving greater network capacity through improved spectral efficiency. Services include wide-area wireless voice telephony, video calls, and broadband wireless data, all in a mobile environment. Additional features also include HSPA (High Speed Packet Access) data transmission capabilities able to deliver speeds up to 14.4 Mbps on the downlink and 5.8 Mbps on the uplink.

The first commercial 3G network was launched by NTT DoCoMo in Japan branded FOMA, based on W-CDMA technology on October 1, 2001 [8]. The second network to go commercially live was by SK Telecom in South Korea on the 1xEV-DO (Evolution- Data Optimized) technology in January 2002 followed by another South Korean 3G network was by KTF on EV-DO in May 2002. In Europe, the mass market commercial 3G services were introduced starting in March 2003 by 3 (Part of Hutchison Whampoa) in the UK and Italy. This was based on the W-CDMA technology. The first commercial United States 3G network was by Monet Mobile Networks, on CDMA2000 1x EV-DO technology and the second 3G network operator in the USA was Verizon Wireless in October 2003 also on CDMA2000 1x EVDO.

The first commercial 3G network in southern hemisphere was launched by Hutchison Telecommunications branded as Three using UMTS in April 2003. The first commercial launch of 3G in Africa was by EMTel in Mauritius on the W-CDMA standard. In North Africa (Morocco), a 3G service was provided by the new company Wana in late March 2006. Roll-out of 3G networks was delayed in some countries by the enormous costs of additional spectrum licensing fees. In many countries, 3G networks do not use the same radio frequencies as 2G, so mobile operators must build entirely new networks and license entirely new frequencies; an exception is the United States where carriers operate 3G service in the same frequencies as other services. The license fees in some European countries were particularly high, bolstered by government

auctions of a limited number of licenses and sealed bid auctions, and initial excitement over 3G's potential. Other delays were due to the expenses of upgrading equipment for the new systems. Still several major countries such as Indonesia have not awarded 3G licenses and customers await 3G services. China delayed its decisions on 3G for many years. In January 2009, China launched 3G but interestingly three major companies in China got license to operate the 3G network on different standards, China Mobile for TD-SCDMA, China Unicom for WCDMA and China Telecom for CDMA2000 [3].

Fourth Generation (All-IP)

The emergence of new technologies in the mobile communication systems and also the ever increasing growth of user demand have triggered researchers and industries to come up with a comprehensive manifestation of the up-coming fourth generation (4G) mobile communication system. In contrast to 3G, the new 4G framework to be established will try to accomplish new levels of user experience and multi-service capacity by also integrating all the mobile technologies that exist (e.g. GSM - Global System for Mobile Communications, GPRS - General Packet Radio Service, IMT-2000 - International Mobile Communications, Wi-Fi - Wireless Fidelity, Bluetooth) (see Fig. 1) [4].

The fundamental reason for the transition to the All-IP is to have a common platform for all the technologies that have been developed so far, and to harmonize with user expectations of the many services to be provided. The fundamental difference between the GSM/3G and All-IP is that the functionality of the RNC and BSC is now distributed to the BTS and a set of servers and gateways. This means that this network will be less expensive and data transfer will be much faster [2]. 4G will make sure - "The user has freedom and flexibility to select any desired service with reasonable QoS and affordable price, anytime, anywhere." 4G mobile communication services started in 2010 but will become mass market in about 2014-15.

IMT-Advanced 4G standards will usher in a new era of mobile broadband communications, according to the ITU-R. IMT Advanced provides a global platform on which to build next generations of interactive mobile services that will provide faster data access, enhanced roaming capabilities, unified messaging and broadband multimedia. According to ITU, "ICTs and broadband networks

have become vital national infrastructure — similar to transport, energy and water networks — but with an impact that promises to be even more powerful and far-reaching. These key enhancements in wireless broadband can drive social and economic development, and accelerate progress towards achieving the United Nations’ Millennium Development Goals, or MDGs.” [12]. The current agreements on the requirements for IMT-Advanced are:

- Peak data rate of 1 Gbps for downlink (DL) and 500 Mbps for uplink (UL).
- Regarding latency, in the Control plane the transition time from Idle to Connected should be lower than 100ms. In the active state, a dormant user should take less than 10ms to get synchronized and the scheduler should reduce the User plane latency at maximum.
- Downlink peak spectral efficiency up to 15 bps/Hz and uplink peak spectral efficiency of 6.75 bps/Hz with an antenna configuration of 4×4 or less in DL and 2×4 or less in UL.
- The average user spectral efficiency in DL (with inter-site distance of 500m and pedestrian users) must be 2.2 bps/ Hz/cell with MIMO 4×2 , whereas in UL the target average spectral efficiency is 1.4 bps/Hz/cell with MIMO 2×4 .
- In the same scenario with 10 users, cell edge user spectral efficiency will be 0.06 in DL 4×2 . In the UL, this cell edge user spectral efficiency must be 0.03 with MIMO 2×4 .
- Mobility up to 350 km/h in IMT-Advanced.
- IMT-Advanced system will support scalable bandwidth and spectrum aggregation with transmission bandwidths more than 40MHz in DL and UL.
- Backward compatibility and inter-working with legacy systems.

After completion of its Release-8 specifications, Third Generation Partnership Project (3GPP) has already planned for a work item called LTE-Advanced to meet the IMT-Advanced requirements for 4G. Also, WiMAX Forum and IEEE are also evolving WiMAX through IEEE 802.16m or WiMAX-m to satisfy 4G requirements. Table 1 summarizes the generations of wireless technology.

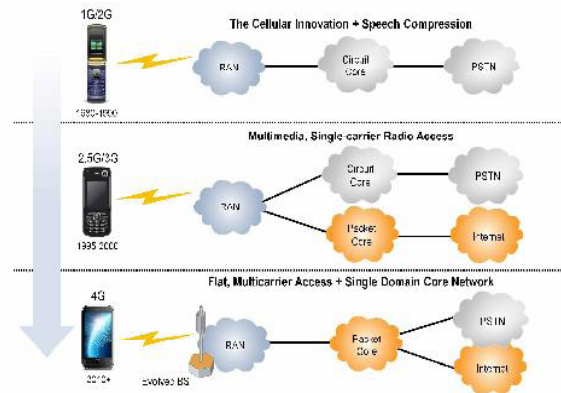


Fig. 1. Mobile Network Evolution.

GPP Evolution

3GPP, the body which defined the LTE specifications, has a well established evolutionary scheme that is likely to continue for some time. Release 99 (R99) defined the original dual-domain UMTS system that supports both circuit-switched voice services and packet switched access. Release 4 (R4) saw the earliest phase of network IP adoption with the deployment of a bearer-independent circuit-switched architecture that disassociated the telephone switches into media gateways and controllers (soft switches).

Releases 5 through 7 are dominated by techniques to increase the spectral efficiency, and thereby extend the viability of the limited underlying W-CDMA technology. The resulting HSPA+ (High Speed Packet Access, with enhancements) can theoretically achieve peak data rates of 42 Mbps under ideal conditions. In addition to improving spectral efficiency, R5 specifies the initial design of IP Multimedia Subsystem (IMS) – an IP services environment.

IP began to proliferate throughout the network with the evolution of interfaces originally delivered on ATM/E1 now being migrated to IP for cost and efficiency purposes.

Release 8 (R8) defines the Long Term Evolution (LTE) system as a break with the past. It marks the start the transition to 4G technologies and networks.

Release 9 (R9) offers enhancements to LTE, including definition of Home eNodeBs for improved residential and in-building coverage.

While first adopters of LTE are working to develop and deploy R8-based LTE systems, work is underway on defining still more improvements to LTE. In particular, the yet to be completed Release 10 (R10) recommendation that defines LTE-Advanced, is a full-featured 4G system which includes 8x8 MIMO, channel aggregation up to 100 MHz, and relay repeaters. It additionally seeks to improve operational efficiencies by supporting a range of self optimizing, self healing capabilities that enable the network to execute tasks that in earlier technologies have been carried out manually[5].

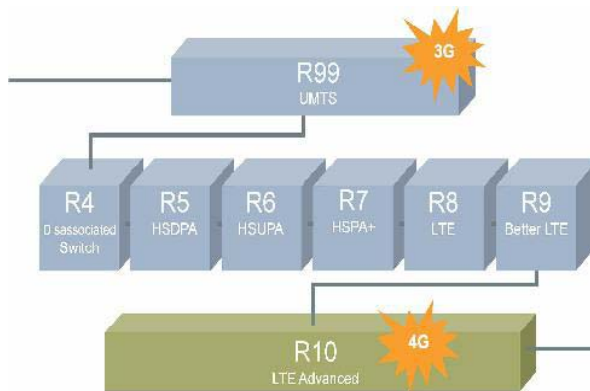


Fig.2. 3GPP Roadmap.

LTE Architecture

Figure 3 illustrates the overall LTE architecture, which is marked by the elimination of the circuit-switched domain and a simplified access network. The functional entities depicted in Figure 3 can be physically co-located, or reside in dedicated hardware according to the network operator’s needs.

The LTE system is comprised of two networks: the E-UTRAN and the Evolved Packet Core (EPC). The result is a system characterized by its simplicity, a non-hierarchical structure for increased scalability and efficiency, and a design optimized to support real-time IP-based services. The access network, E-UTRAN is characterized by a network of Evolved-NodeBs (eNBs), which support OFDMA and advanced antenna techniques. Each eNB has an IP address, and is part of the all-IP network. Dedicated radio network controllers, which were present in earlier

generation access networks, are not required; the eNBs in LTE collaborate to perform functions such as handover and interference mitigation.

Similarly, the all-IP packet core network enables the deployment and efficient delivery of packet-oriented multimedia services, through the IP Multimedia Subsystem (IMS). This results in lower costs and rapid deployment of new services for network operators.

While there are recognizable parallels to the 3GPP UMTS packet core network, the EPC is a significant departure from the legacy packet core which enables growth in packet traffic, higher data rates, and lower latency, and support for interworking with several wireless access technologies.

From 3G to 4G

The demand for ever-higher data rates, higher capacity, higher user throughput, lower latency, more efficient use of the radio spectrum, and more flexibility fuelled the need for a departure from the inherent limitations of UMTS (and its many evolutions) by, among other things, its CDMA-based air interface. LTE ensures a viable future for mobile broadband by enabling:

1. data rates an order of magnitude higher than single carrier spread spectrum radio can provide (over 300 Mbps downlink and 50 Mbps uplink, compared to 14 Mbps uplink and 5.76 downlink on UMTS HSPA).
2. reduced transit times for user packets (reduced latency), an order of magnitude shorter than can be provided in hierarchical 3G networks (5 ms for data and under 100 ms for signaling).
3. the ability for strict Quality of Service (QoS) control of user data flows with the possibility of these being coupled with various charging schemes.

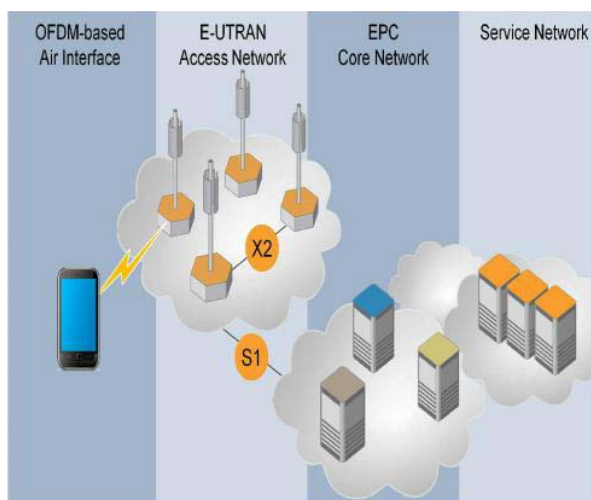


Fig.3. LTE Architecture

The EPC Architecture

LTE's packet domain is called the Evolved Packet Core (EPC). It is a flat all-IP system designed for:

- _ much higher packet data rates _ significantly lower-latency
 - _ the ability to optimize packet flows within all kinds of operational scenarios having to do with bandwidth rationing and charging schemes.
 - _ explicit support for multiple radio access technologies in the interests of seamless mobility, and
 - _ greater system capacity and performance
- Six nodes are defined to meet these goals:

MME: Somewhat analogous to the distribution of control and bearer data of the MSC found in 3GPP R4, LTE separates control from bearer in the design of the EPC.

The Mobility Management Entity (MME), which supports many functions for managing mobiles and their sessions, also controls establishment of EPS bearers in the selected gateways.

S-GW: The Serving Gateway (S-GW) is responsible for anchoring the user plane for inter-eNB handover and inter-3GPP mobility. An S-GW functionally resembles a 3G SGSN without the mobility and session control features. It routes data packets between the P-GW and the E-UTRAN.

P-GW: The Packet Data Network Gateway (P-GW) acts as a default router for the UE, and is responsible for anchoring the user plane for mobility between some 3GPP access systems and all non-3GPP access systems.

HSS: The Home Subscriber Server is the master data base that stores subscription-related information to support call control and session management entities.

PCRF: The Policy and Charging Control Function (PCRF) is the single point of policy-based QoS control in the network. It is responsible for formulating policy rules from the technical details of Service Data Flows (SDF) that will apply to a user's services, and then passing these rules to the P-GW for enforcement.

ePDG: The evolved Packet Data Gateway (ePDG) is used for interworking with un-trusted non-3GPP IP access systems.

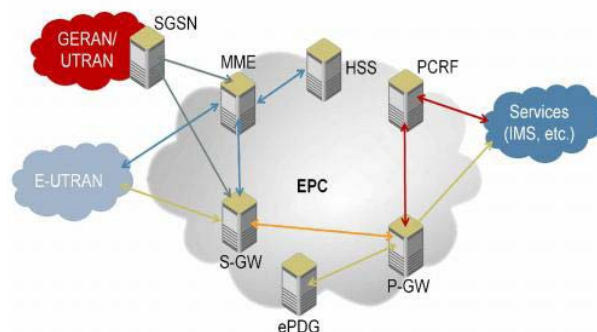


Fig.3. Key EPC Entities.

E-UTRAN

The Evolved UMTS Terrestrial Radio Access Network (EUTRAN) implements the LTE access network as a network of eNBs. A key difference between UTRAN and E-UTRAN is the absence of a centralized radio network controller.

What were once centrally-coordinated functions are now distributed into the eNBs and the X2 interfaces between them. The eNB is responsible for many functions including:

- _ Radio Resource Management;
 - _ IP header compression and user data encryption;
 - _ the scheduling and allocation of uplink and downlink radio resources, and
 - _ coordinating handover with neighboring eNBs.
- eNBs can communicate with multiple gateways for load sharing and redundancy.

IMS

3GPP has developed a complete service network system for mobile networks, called the IP Multimedia Subsystem (IMS). It is a complete, SIP-based control architecture that includes charging, billing and bandwidth management. As

such, it defines its own formal interfaces with the IETF for any protocol extensions.

IMS is intended to occupy the core of tomorrow's converged networks, and is likely to be the chief enabler of accelerated network convergence with the promise of flexible service delivery. Mobile operators will count on LTE to

Conclusion

The last few years have witnessed a phenomenal growth in the wireless industry. The ever increasing demands of users have triggered researchers and industries to come up with a comprehensive manifestation of the up-coming fourth generation (4G) mobile communication system. As the history of mobile communications shows, attempts have been made to reduce a number of Technologies to a single global standard.

The first generation (1G) has fulfilled the basic mobile voice, while the second generation (2G) has introduced capacity and coverage. This is followed by the third generation (3G), which has quest for data at higher speeds to open the gates for truly "mobile broadband" experience, which will be further realized by the fourth generation (4G).

LTE is strongly positioned to lead the evolution in the communications industry for several years. It improves spectral efficiency, simplifies deployment of all-IP real-time services, facilitates integration with non-wireless networks, and supports interworking with legacy wireless technologies. It achieves all of these things through a flat, scalable architecture that is designed to manage and maintain service QoS in a mobile environment with significantly higher data throughput.

Implementing IMS with LTE enables acceleration network convergence with the promise of flexible service delivery. Mobile operators will count on LTE to implement cost effective network changes as preparation for IMS.

Deploying SIP-based control architectures on broadband wireless IP-based LTE is a natural fit for IMS. For operators overlaying existing wireless networks, LTE supports interworking with the legacy 3GPP and non-3GPP wireless

implement cost-effective network changes as preparation for IMS. Deploying SIP-based control architectures on broadband wireless IP-based LTE naturally implies services ought to reside in the IMS.

accesses, for service continuity that is transparent to the access technology.

To reap the benefits of LTE, network operators will have to face new challenges. Operators must consider migration strategies from legacy 2G/3G networks, reconsider how services are developed, and deploy IP networks which can deliver low latency end-to-end in order to support realtime QoS. This may include interim device strategies, leasing access on 3rd-party IP networks, implementing IMS architectures, and supporting real-time services like VoIP and streaming video over IP. Operators wanting to implement real-time services like VoIP will have to carefully monitor network latency. As these challenges are met and LTE is deployed, operators will recognize significant overall cost savings across the network and significant future revenue opportunities.

Reference:

- [1]. ITU (2009). "Measuring the Information Society; The ICT Development Index", [Online] Available: http://www.itu.int/ITU-D/ict/publications/idi/2009/material/IDI2009_w5.pdf
- [2]. 3gamericas (2010). "Transition to 4G: 3GPP Broadband Evolution to IMT-Advanced", Rysavy Research/3G Americas. [Online] Available: www.rysavy.com/PR/3GA_PR_2010_09.pdf
- [3]. K. Mishra, Ajay. "Fundamentals of Cellular Network Planning and Optimization, 2G/2.5G/3G...Evolution of 4G", John Wiley and Sons, 2004.
- [4]. Pereira, Vasco & Sousa, Tiago. "Evolution of Mobile Communications: from 1G to 4G", Department of Informatics Engineering of the University of Coimbra, Portugal 2004.
- [5]. www.tektronixcommunications.com/LTE.

Conceptual Discrete Wavelet Transformation Speech Hashing for Content Authentication

Mahdi Nouri¹, Nooshin Farhangian²

Department of Electrical Engineering
Ghiasodin Institute of Higher Education
Abeyk, Iran

mnouri@mtu.edu¹, n.farhangian@gmail.com²

Zahra Zeinolabedini³, Nasim Fekri⁴

Department of Electrical Engineering
Ghiasodin Institute of Higher Education
Abeyk, Iran

z.zeinolabedini@gmail.com³, n67.fekri@gmail.com⁴

Abstract— Conceptual hash functions provide a tool for fast and reliable identification of content authentication. Robust hashing for multimedia authentication is an emergent research area. A different key-dependent robust speech hashing based upon speech construction model is proposed in this article. The proposed hash function is based on the essential frequency series. Robust hash is calculated based on linear spectrum frequencies which model the verbal territory. The correlation between LSFs is decoupled by discrete wavelet transformation (DWT). A randomization structure controlled by a secret key is used in hash generation for random feature selection. The hash function is key-dependent and collision resistant. Temporarily, it is extremely robust to content protective operations besides having high accuracy of tampering localization. They are found, the first, to perform very adequately in identification and verification tests, and the second, to be very robust to a large range of attacks. Furthermore, it can be addressed the issue of security of hashes and proposed a keying technique, and thereby a key-dependent hash function.

Keywords—component; DWT; Content-based authentication; robust hashing; least-square periodicity

I. INTRODUCTION

A conceptual speech hash function offers a tool for firm and reliable identification of content. A new audio hash functions is presented based upon précis of the time-frequency spectral characteristics of an audio file. In this work, the procedure is developed for summarizing a stretched audio signal into a short signature sequence, which can then be utilized to recognize the original record. We call this signature the conceptual hash function, because it is sensed to reveal the perceptible component of the content. In other words, the main encourage is obtaining audio hash functions which are insensitive to “reasonable” signal processing and edition process, for example compression, filtering, conversion of sampling rate and so forth, thus so sensitive to any changes in content. Conceptual hash functions can be used as an instrument to search for an exact record in a databank, to sense content tampering attacks, and so forth [1]. This kind of hash functions map speech to a short binary string based on the speech’s conceptual properties, is proposed as a different solution for automatic speech keying and speech content authentication. Contrasting to the traditional cryptographic hash function, which is exceedingly sensitive to the input data, the speech hash function permits for some changes of the speech whereas distinguishing all speech parts from another.

Generally, the speech hash function requires two properties: discrimination, which means that conceptually dissimilar speech parts must have different hash vectors, and conceptual robustness, which means that identical speech parts should have the similar hash vector.

Because of wide application in automatic audio keying and speech authentication, the conceptual speech hash function has been widely studied recently [2-5]. However, there are few speech hash functions available. A compressed domain speech hash structure participated with a mixed excitation linear prediction codec is proposed in [6]. It applies partial bits of the speech bit stream with the linear spectral frequencies; the hash vector round based upon NMF of linear prediction coefficients is proposed [7]. The two features of the conceptual hash function which are so imperative, uniqueness and robustness. The uniqueness condition which is sometimes called randomness, infers that the hash sequence should reveal the content of the audio file in a unique manner.

There exist a number of audio hashing procedures in the literature. Mihcak and Venkatesan [9] extract statistical factors from arbitrarily particular regions of the time-frequency demonstration of the signal; Haitsma et al. offered an audio hashing algorithm [8], where the hash extraction system is based upon measuring the threshold of the energy changes between frequency bands, in another algorithm. In this work, one conceptual audio hashing procedure is investigated that operate in the frequency domain, and use the inherent periodicity of audio signals. In this paper, an algorithm based upon the Dyadic Wavelet Transform has been inspected for detection of musical signals. Wavelet transform is based on the idea of filtering a signal $f(t)$ with a translated and dilated versions of a archetype function $Y(t)$. This function is named the mother wavelet and it has to gratify certain requirements [11] Dyadic Wavelet Transform (DWT), is the special case of CWT. when the scale parameter is discretized along the dyadic grid properly chosen wavelet, the wavelet transform modulus most denote the points of sharp variations of the signal [12-14]. This property of DWT has been established very convenient for detecting periods of speech signals [15]. It can be conjectured that the component of DWT periodicity profile of an audio frame can be used as a signature for tamper control and identification. The periodicity property of the audio signals has been used in such applications as voice activity detection silence detection, and speech compression [10]. This method is

correlation based. This correlation-based Dyadic Wavelet hashing, called CDWH.

II. PERIODICITY-BASED HASH FUNCTIONS

A. Periodicity measure by a correlation-based analysis

The first peak of the autocorrelation of the linear prediction residue indicates the pitch period and is commonly used as a pitch estimator. This correlation-based periodicity estimate, called CPE, has the following expression:

$$\widehat{P}_0 = \begin{cases} \arg \max R(K), \text{ for } K \neq 0 \text{ if } R(\widehat{P}_0) \geq 0.5 \\ 0 \text{ if } R(\widehat{P}_0) < 0.5 \end{cases} \quad (1)$$

The efficacy of the CPE method is enhanced by a four-tap prediction and decimation process. The advantage of the correlation-based method is that it requires about three times less computation as compared to the parametric estimation.

B. The simulated attacks

The template is used to format your paper and style the text. All margins, column widths, line spaces, and text fonts are prescribed; please do not alter them. You may note peculiarities. For example, the head margin in this template measures proportionately more than is customary. This measurement and others are deliberate, using specifications that anticipate your paper as one part of the entire proceedings, and not as an independent document. Please do not revise any of the current designations.

In this section, we focus on transform-domain hash functions in contrast to the previous section, where we essentially worked on the time domain. More specifically, the audio signal is divided into possibly overlapping frames and each frame is represented by its mel-frequency cepstral coefficients (MFCCs), which are short-term spectral-based features [15]. A singular value decomposition (SVD) further summarizes these features. Note that in the SVD-based method we use the original signal, and not its low pass filtered version, as in the periodicity-based schemes.

$$R = \frac{N \sum_f X(f)Y(f) - \sum_f X(f) \sum_f Y(f)}{\sqrt{[N \sum_f X^2(f) - (\sum_f X(f))^2][N \sum_f Y^2(f) - (\sum_f Y(f))^2]}} \quad (2)$$

$$d = \frac{1}{N} \sqrt{\sum_f (X(f) - Y(f))^2} \quad (3)$$

III. PROPOSED ROBUST SPEECH HASHING

The proposed scheme has two phases: feature extraction and hash modeling. In the feature extraction phase, perceptual features are extracted from speech signals such that the hash values could be robust to content-preserving operations. In the following hash modeling phase, a random secret key is employed to generate the secure hash sequence. Here, the extracted feature vectors are compressed to be compact and

coded with the secret key to be randomized. The whole process is detailed in the following subsections.

A. Feature Extraction

Linear prediction coding (LPC) is widely used for speech coding and recognition, in which voice is modeled as the response of a vocal tract filter to a glottal excitation [9]. LPC coefficients model the slowly varying transfer function of the vocal tract. The varying transfer function determines the vowels which are important to speech perception. As a linear spectral representation of LPC coefficients, linear spectral frequencies (LSFs) have been widely used in speech coding and other speech processing applications. Therefore, as the content-based robust speech feature, LSFs are employed to generate the hash value in this letter.

To verify the discriminative and robust nature of the proposed hash function, it was applied to 1,000 speech clips (16 bits signed, 8 kHz) with various contents. On the waveform, the X-axis represents time, with past to future moving from left to right. The Y-axis represents intensity of the sound. The waveform exactly reflects the nature of a sound, which is just a series of fluctuations across time. Note that the signal is essentially periodic and repeating, though it has a somewhat random character. In fact, most sounds can be simply thought of as a combination of different repeating signals with various amplitudes and frequencies. Essentially, if an individual spectrum is thought of as a cross-section of a mountain, for example, then a spectrogram corresponds to a topographical map, composed of many spectra arranged side by side on their ends. A graph of a spectrum is produced by taking a small area around a single point and determining the amplitude and frequency of the signals immediately surrounding that point. They are then plotted on to the graph above, where the X-axis is frequency in Hz, and the Y-axis is amplitude in decibels.

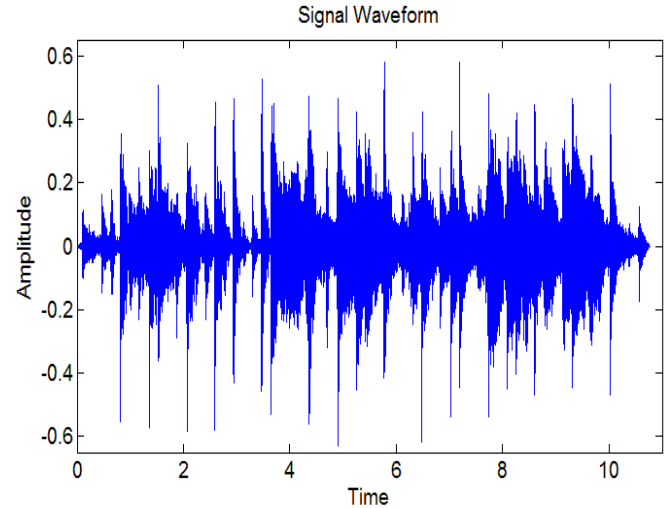


Figure 1. intensity of the sound per time of main audio file

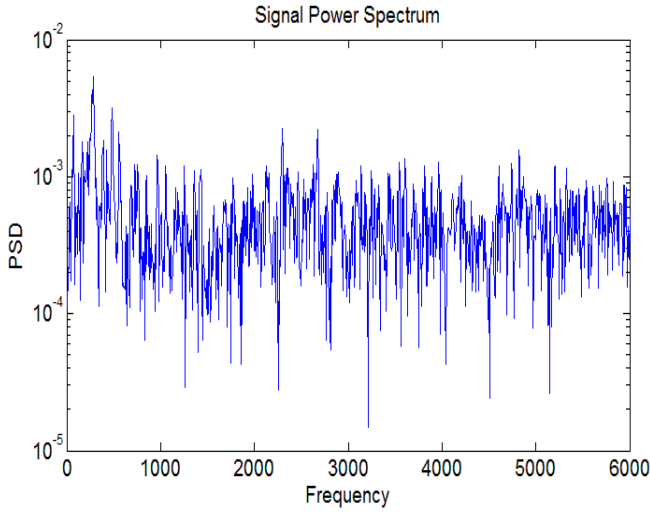


Figure 2. Power spectrum density per frequency of main audio file

Now it is easy to determine the frequency of waves of any amplitude, and vice versa, which would have been very difficult to determine from the waveform above. Bear in mind, however, that this spectrum is only of a single point, meaning that an entire speech sound cannot be easily analyzed using a spectrum.

Therefore, a random matrix of DWT coefficients is obtained. There are many alternative transformations available, such as DCT, FFT and PCA. Because of the good energy compaction property, DWT is used to decorrelate the LSFs and extract new independent features in hash modeling.

B. Units

A circle of radius "b" rolls on the outside of a circle of radius "a". "P" is the point on the b-circle of initial contact with the a-circle. As the b-circle rolls the point "P" traces out a curve in the plane. This is an epicycloid. There will be "a/b" returns contacts of "P" with the a-circle as the b-circle rolls. So, when $a/b = N$ is an integer, we get a closed figure with N vertices (in one traversal of the a-circle).

The shape of the epicycloid is totally determined by the single number N. The general two-dimensional chaotic nonlinear map in the algorithm is based on the parametric form of an epicycloid is:

$$\begin{cases} x(\theta) = (a + b)\cos\theta - b\cos\left(\frac{(a + b)}{b}\theta\right) \\ y(\theta) = (a + b)\sin\theta - b\cos\left(\frac{(a + b)}{b}\theta\right) \end{cases} \quad (4)$$

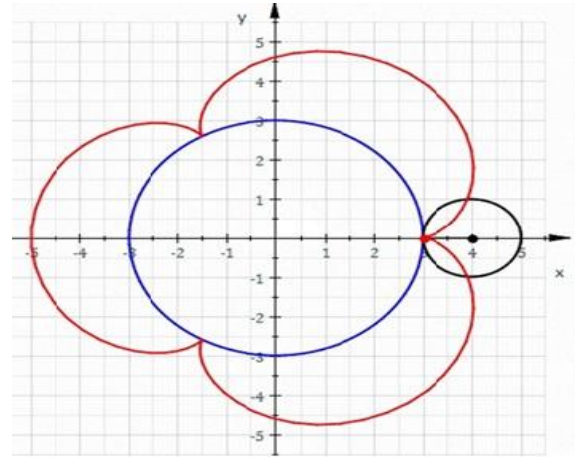


Figure 3. an epicycloid graph

C. Hash Modeling

Security is essential when robust hashing is applied for content authentication. A keyed randomization scheme is introduced in hash modeling to guarantee that an unauthorized user cannot forge a valid hash without the key. The hash modeling phase contains six steps.

- 1) Seeded with a secret key, one random number sequences I_T is generated by a uniform pseudo-random number generator (PRNG). They are used to select blocks from the DWT coefficient matrix (L).
- 2) From frame length, a pair of blocks $r_{i,1}$ and $r_{i,2}$ are chosen with the same size. $I_T(i, 1)$ and $I_T(i, 2)$ are the frame IDs of the first frame in $r_{i,1}$ and $r_{i,2}$, respectively.
- 3) $r_{i,1}$ and $r_{i,2}$ are transformed by a two-dimensional DWT components. DWT coefficients are taken with respect to robustness. The resulting features are denoted as $d_{i,1}$ and $d_{i,2}$. $d_{i,1}$ used when (i) is even and $d_{i,2}$ used when (i) is odd

$$\begin{cases} d_{i,1} = (r_{i,1} + r_{i,2}) \cos(r_{i,1} * r_{i,end}) - r_{i,2} \cos\left(\frac{(r_{i,1} + r_{i,2})}{r_{i,2}}(r_{i,1} * r_{i,end})\right) \\ d_{i,2} = (r_{i,1} + r_{i,2}) \sin(r_{i,1} * r_{i,end}) - r_{i,2} \sin\left(\frac{(r_{i,1} + r_{i,2})}{r_{i,2}}(r_{i,1} * r_{i,end})\right) \end{cases}$$

- 4) The next input rotation matrix is given below by rotating component.

$$r(j + 2, :) = d(1, :)$$

- 5) The hash components, denoted as $h_i(j)$ is decided by the sign of feature $d_{i,1}$ and $d_{i,2}$ difference of last rotation, as given in (1), where $j \in [1, F]$

$$h(i) = d_{i,1} - d_{i,2}$$

- 6) The steps from 2 to 4 are repeated for t_i times ($i \in [1, 64]$) Thus, the final hash sequence is a $F \times t$ binary matrix, which is the content-based signature of a one second speech clip.

Besides security, random feature extraction has another advantage over the method proposed in [10]. As the pairs for comparison are randomly selected, both global and local features are used to enhance the sensitivity to tampering. The bitrate of CDWH is 0.512 Kbps, which is lower than that of rMac (5.5 KHz) [5] and of SCA (0.9 KHz) [10]. With high performance speech coders, raw speech samples could be generated with low cost. Therefore, CDWH of speech signal has versatility as well as low complexity.

IV. EXPERIMENTAL RESULTS

Simulation experiments have been performed in order to test, (i) the robustness of the perceptual hash for identification, where the critical behavior is the statistical spread of the hash function when an audio document is subjected to various signal processing attacks; (ii) the uniqueness of the perceptual hash, where the important behavior is the fact that the hashes differ significantly between two different contents. In other words, in the first case, we want to identify a document and its variants under signal processing attacks. In the second case, we documents with different contents should be classified, so that if a document is wanted to verify, the others in the database appear as “impostors.”

A. Discrimination

BERs between 5000000 pairs of hashes are calculated. The hashes are extracted from speech excerpts of different contents. The speech excerpts are randomly selected from the database described above. The normal probability plot of the measured BERs is shown in Fig. 5. If and only if the bits are independent and identically distributed (the probability of 0/1 is equal to 0.5), the bit error rate (BER) of the two hash sequences is normally distributed, with a mean of p and a standard deviation of $\sigma_0 = \sqrt{p \times (1-p)/N} = \sqrt{1/4N}$ (N is the number of bits in a hash sequence) [8]. With the parameter in proposed scheme, the mean and standard deviation of normal distribution should be 0.5 and 0.0221, respectively. From experimental results, we get the mean and the standard deviation, which are very close to the normal distribution values. Hence, hash sequence is random and collision resistant. When the threshold is set to 0.1, the false positive rate is 3.4655×10^{-20} , which is much lower than the false positive rate in [10]. For each speech clip, the distance of its hash vector and the hash vector of each of the remaining 2879 speech clips was calculated by Eq.15, and 497,121 distance values were obtained. A comparison between the probability density distribution of these distance values and the normal distribution shown in Fig. 5 indicates that the distance value has an approximately normal distribution. The expected value and standard deviation are $\mu = 0.49998$ and $\sigma = 0.00912$, respectively.

The FAR (calculated by (15)) varying with the threshold is shown in Table II. Fig. 5 shows a comparison of the FAR curve obtained by theoretical analysis (see Fig.5) with that obtained by the experimental values. The probability density distribution of the distance values follows the normal approximation fairly well; thus, it is verified that the threshold obtained by Fig.5 can be used in practice with reasonable accuracy.

B. Robustness

Speech excerpts are subjected to the following content preserving manipulations.

Table II lists the average value of the BERs. Now it's time to discuss the robust performance of proposed scheme.

First, the BER caused by transcoding errors is a little bit higher than by resampling. Second, the resulting BER is below the pre-specified threshold 0.25, that is, our CDWH is robust to noise addition.

Third, there is no difference between the LSFs of the original speech and those after volume change. It is due to the reason that volume amplifying and reducing do not change the vowels. Fourth, since MP3 compression is based on non-linear psychoacoustic model, MP3 re-encoding delivers higher BERs than speech transcoding. Fifth, the BER is still small (well below 0.1) even when the portion of the cropped speech reaches 40%. It is because of global features are used to generate the hash in CDWH. Finally, when the frame desynchronization is large, time scaling usually delivers large BERs. That is why the proposed CDWH is breakable when the time scaling is larger than 2%.

C. Statistic analysis of diffusion and confusion

Confusion and diffusion are two essential design criteria for encryption algorithms, including hash functions. Shannon initiated diffusion and confusion in order to conceal message redundancy [18,19]. Hash function, like encryption system, requires the plaintext to diffuse its effects into the whole Hash space. This means that the correlation between the message and the corresponding Hash value should be as small as possible. Diffusion means spreading out the influence of a single plaintext symbol over many audio symbols so as concealing the statistical structure of the audio file. Confusion means the utilizing of transformations to make difficult the dependence of audio file statistics. In the hash value in audio pairs format each audio symbol can be changed. Therefore, the perfect diffusion effect should be that any minute change in the initial condition leads to a 50% changes in average of the all symbols, change probability of each symbol. Regularly six statistics are defined as follows:

Minimum changed audio symbol number:

$$B_{min} = \min(\{B_i\}_1^N) \quad (5)$$

Maximum changed audio symbol number:

$$B_{max} = \max(\{B_i\}_1^N) \quad (6)$$

Mean changed audio symbol number:

$$\bar{B} = \frac{1}{N} \sum_1^N B_i \quad (7)$$

Mean changed probability:

$$P = \frac{\bar{B}}{L} \times 100 \quad (8)$$

Standard variance of the changed audio symbol number:

$$\Delta B = \sqrt{\frac{1}{N-1} \sum_{i=1}^N (B_i - \bar{B})^2} \quad (9)$$

Standard variance:

$$\Delta P = \sqrt{\frac{1}{N-1} \sum_{i=1}^N \left(\frac{B_i}{L} - P\right)^2} \times 100 \quad (10)$$

Where N is the total number of tests and B_i is the amount of changed audio symbols in the i_{th} test.

The advantage of the correlation-based method is that it requires about three times less computation as compared to the parametric estimation.

$$R(K) = \frac{\left(\frac{1}{N-K}\right) \sum_{i=0}^{N-K} s(i)s(i+K)}{\left(\frac{1}{N}\right) \sum_{i=0}^N s^2(i)} \quad (11)$$

TABLE I
STATISTICAL PERFORMANCE OF THE PROPOSED ALGORITHM

Frame	600	800	1000	1200	1400	1600
B	0.0744	0.0858	0.0874	0.0880	0.0507	0.0487
ΔB	0.1278	0.1481	0.1545	0.1599	0.0987	0.1003
P	51.937	48.997	0.0508	50.153	51.122	50.455
ΔP	0.0445	0.0689	0.0898	0.1118	0.0802	0.0937
R	0.8995	0.8838	0.8795	0.9050	0.8905	0.8979

D. Security aspects of the audio hash functions

The security of the hash extraction becomes important in audio authentication schemes. One common way to provide hash security is to devise a key-based scheme such that for two different keys, K_1 and K_2 , the resulting hash functions become totally independent. Thus we minimize the probability of collision, that is, we want to guarantee that two distinct inputs yield different hash functions and that the hash sequences are mutually independent. Notice that secure fingerprinting requires that the pirate should not be capable of extracting the hash value of the content without knowledge of some secret key. This would, for example, allow him to change the content while preserving the hash, that is, find a collision which would circumvent any hash-based authentication mechanism being used. As another example, it could also enable him to manipulate the bits while preserving the content and yet change the hash. This would be done, for example, when a pirate may want to avoid being detected by a copyright controller for unauthorized use of some content. One way to arrive at a key-based hash function is to project the resulting hash sequences onto key-dependent random bases. Another scheme would be to subject the analog hash sequence to random quantization [20]. In this scheme, the hash sequence is quantized using a randomized quantizer, and the quantizer itself becomes the source of randomness in the hash function's output. A third scheme could be based on random permutation of the

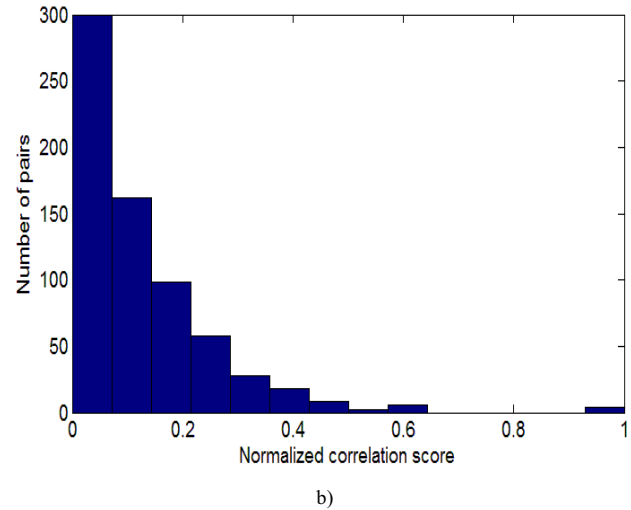
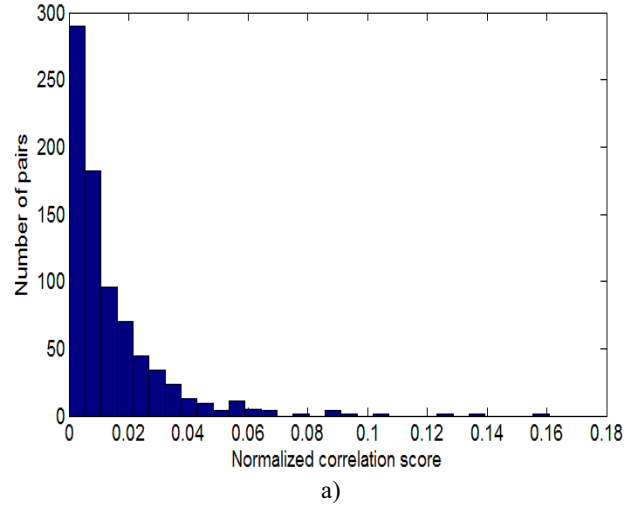


Figure 4. Histograms of the difference of the hash functions with 1000 speech records: (a) hashes of the different objects (solid line) and those of the attacked versions of the same object (dashed line); (b) hashes obtained from the same object with different keys.

observation frames with possible overlaps. Thus we generate a key-based sequence of visiting positions and translate in saccades the frame window according to this sequence (recall that we used 25-millisecond windows with 50% overlap).

E. Hash Matching

The problem of hash matching can be formulated as the hypothesis testing using the hash function $H(\cdot)$ and the distance measure $D(\cdot, \cdot)$.

L_0 : Two speech clips $\mathbf{a}_1, \mathbf{a}_2$ are from the same speech if

$$D(H(\mathbf{a}_1), H(\mathbf{a}_2)) < \tau \quad (12)$$

L_1 : Two speech clips $\mathbf{a}_1, \mathbf{a}_2$ are from different speech if

$$D(H(a_1), H(a_2)) \geq \tau \quad (13)$$

Where τ is a predetermined threshold, which can be obtained for a given false accept rate (FAR). FAR, denoted by R_{FA} , is the probability that L_0 is accepted when L_1 is true. In the proposed scheme, the square of the Euclidean distance is utilized to measure the distance between any two hash vectors h_1 and h_2 .

$$x = D(h_1, h_2) = \frac{1}{L_h} \sum_{n=1}^{L_h} [h_1(n) - h_2(n)]^2 \quad (14)$$

Where L_h is the length of the hash vector. By the central limit theorem, the above distance measure has a normal distribution if L_h is sufficiently large and the contributions in the sums are sufficiently independent. Assuming that the distance measure can be approximated as the normal distribution $N(\mu, \sigma)$, the FAR is given as

$$R_{FA} = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\tau} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right] dx \quad (15)$$

Then, for a given R_{FA} , the threshold τ can be determined by (15), theoretically.

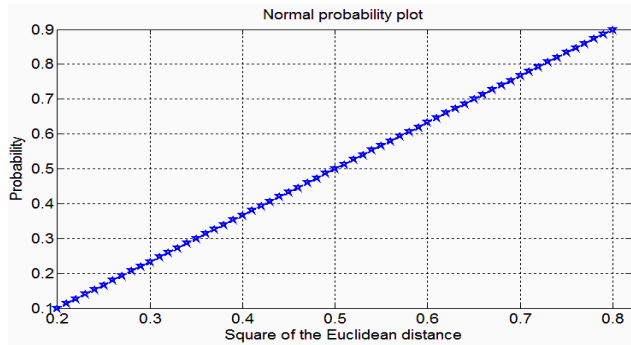


Figure 5. Comparison of probability density distribution of the distance values plotted as '*' and normal distribution.

TABLE II
FAR varying with threshold

τ	R_{FA}
0.1	3.4655×10^{20}
0.2	5.0479×10^{15}
0.3	2.3497×10^{12}
0.4	1.0540×10^9
0.5	1.5428×10^6

V. CONCLUSION

In this paper, a novel keyed robust speech hashing CDWH is proposed. The linear spectrum frequencies are implemented for hash generation. Discrete wavelet transformation is introduced to decorrelate the LSFs and enhance the discriminative capacity. As proposed robust hash function is key-dependent with high collision resistance, it could be used in multimedia authentication system. Experimental results

confirmed that the proposed CDWH is extremely robust against speech noise addition, resampling, transcoding and other modifications. Compared with previous audio hashing, the proposed scheme is increased the tampering localization accuracy from sentence-level to alphabetic-level. For a reliable hash function, the feature extracted should be both discriminative and robust. In this letter, linear prediction analysis and non-negative matrix factorization were investigated for speech hash function. Linear prediction analysis was performed to extract the frequency shaping attributes of the verbal territory to realize the conceptual robustness of the proposed scheme. Experimental outcomes demonstrated the efficiency of the proposed hash function in terms of discernment and conceptual robustness.

ACKNOWLEDGMENT

This paper is supported by Iranian Research Institute for ICT (ITRC).

REFERENCES

- [1] J. S. Seo, J. Haitsma, T. Kalker, and C. D. Yoo, "A robust image fingerprinting system using the Radon transform," *Signal Processing: Image Communication*, vol. 19, no. 4, pp. 325-339, 2004.
- [2] P. Cano et al., "A Review of Audio Fingerprinting," *J. VLSI Signal Process.*, vol. 41, no. 3, 2005, pp. 271-284.
- [3] A. Ramalingam and S. Krishnan, "Gaussian Mixture Modeling of Shorttime Fourier Transform Features for Audio Fingerprinting," *IEEE Trans. Inf. Forensics Security*, vol. 1, no. 4, 2006, pp. 457-463.
- [4] M. Park, H. Kim, and S.H. Yang, "Frequency-Temporal Filtering for a Robust Audio Fingerprinting Scheme in Real-Noise Environments," *ETRI J.*, vol. 28, no. 4, 2006, pp. 509-512.
- [5] Y. Jiao et al., "Key-Dependent Compressed Domain Audio Hashing," *Proc. ISDA*, 2008.
- [6] Y. Jiao, Q. Li, and X. Niu, "Compressed Domain Perceptual Hashing for MELP Coded Speech," *Proc. IHMSp*, 2008, pp. 410-413.
- [7] D.D. Lee and H.S. Seung, "Learning the Parts of Objects by Nonnegative Matrix Factorization," *Nature*, vol. 401, no. 6755, 1999, pp. 788-791.
- [8] T. Kalker, J. Haitsma, and J. Oostveen, "Robust audio hashing for content identification," in *Proc. International Workshop on Content Based Multimedia Indexing (CBMI '01)*, Brescia, Italy, September 2001.
- [9] M. K. Mihcak and R. Venkatesan, "A perceptual audio hashing algorithm: a tool for robust audio identification and information hiding," in *Proc. Information Hiding*, pp. 51-65, Pittsburgh, Pa, USA, April 2001.
- [10] R. Tucker, "Voice activity detection using a periodicity measure," *IEE Proceedings-I: Communications, Speech and Vision*, vol. 139, no. 4, pp. 377-380, 1992.
- [11] Kronland Martinet, R., Morlet, J. and Grossman, A., "Analysis of sound patterns through wavelet transforms", *International Journal of Pattern Recognition and Artificial Intelligence*, 1(2): 273-302, (1987). *Vision*, vol. 139, no. 4, pp. 377-380, 1992.
- [12] Mallat S. and Zhong S., "Characterization of signals from multiscale edges", *IEEE Trans. on Pattern Analysis and Machine intelligence*, 14(7): 710-732, (1992).
- [13] Mallat, S., "Zero-crossing of a wavelet transform", *IEEE Trans. on Information Theory*, 37(4): 1019-1033, (1991).
- [14] Berman, Z. and Baras, J. S., "Properties of the multiscale maxima and zero-crossings representations", *IEEE Trans. on Signal Processing*, 41(12): 3216-3231, (1993).
- [15] Kadambe, S., and Boudreaux-Bartels, G., "Application of the wavelet transform for pitch detection of speech signals", *IEEE Trans. on Information Theory*, 38(2): 917-924, (1992).

Cloud Service for Comprehensive Project Management Software

Ahmad Khan^{#1}, Gundeep Singh Bindra^{#2}, Rohan Arora^{#3},

Nishant Raj^{#4}, Darshan Jain^{#5}, Dhruvad Shrivastava^{#6}

^{#1} MESCOE, Pune, India

^{#2} Department of Computer Science, Columbia University, New York, U.S.A

^{#3,5} SKNCOE, Pune, India

^{#4} HCL Technologies Limited, Noida, India

^{#6} Birla Institute of Technology, Mesra, India

^{#1}ahmed@connectingahmad.info

^{#2}gsb2124@columbia.edu

^{#3}rohan0921@gmail.com

^{#4}nishant.raj@hcl.com

^{#5}darshan.jain13@gmail.com

^{#6}dhru.2882@gmail.com

Abstract— Project Management has long been one of the arduous jobs for individuals as well as companies to pertain to. Yet with the dynamic architecture of Cloud Service facilities, this could easily be managed with respect to the availability and cost efficiency. This paper focuses on the Cloud-Based management of projects through the means of software as a service and its various augmentable utilities. A project management Software is one which is used to incorporate an efficient usage of resources based on the use and utilization, which is best implemented with the help of a Cloud Service. This would in turn simplify the needs of efforts for parties needing to meet software requirements of a project from different third parties and instead, providing it from one source to make the process less complex and highly feasible along with the robustness of Cloud Services.

Keywords— Cloud Services, Project Management, SDLC, Resource utilization, SaaS, Service Oriented Architecture.

INTRODUCTION

For many a people, the software development life cycle (SDLC) tends to be a treacherous task. The primary purpose of SDLC is to manage the entire life cycle of a projects phases, and hence result in a high quality system that meets or exceeds customer expectations, works sustainably and at the same time keeps track of time limitations, maintains efficiently in the current and planned I.T. infrastructure, and is inexpensive to maintain along with cost-effective for enhancement. This, so far has been a hurdle for people to achieve in the best way possible, due to multiple resources from multiple directions. Though, incorporating a model on the Cloud which provides SDLC phases as coordinated services would turn out to reduce the complexities of above mentioned process exponentially. The Cloud model has been underlined emphatically in this paper due to its 'pay-as-you-go' and reasonable requirement aid. Here, components are viewed as services. In the proposed model, services can interact with one another and be providers or consumers of data and behavior, instead of

letting the client gather data from numerous vendors and eventually put it altogether on their own. In Section 3 of this paper we discuss the limitations of current approach. In Section 4 of this paper we present the technique and approach for the implementation of this model.

This research Software will be used to collect the information from the different levels of software development lifecycle shown in Fig. 1 and different people involved in the project/product development. It provides Status and Reports of the different levels of the project/product; it will monitor the entire software development life cycle and provide status and reports at the different levels. It will maintain the details of the client and the developer interaction as all the documentation work will be done by both the client and the software development team who are working on the client project/product. The information updating is done by the authorized person of the company who will be in contact with the client and with the development team basically the project manager can manage all his developers activities and the client can monitor the project managers as well as the developers activities without wasting time and money for traveling. The client who is giving the project/product to be developed to the company can view the documentation that is done. As the client will have a tight control on the different parts of the project/product and also the client will get all the status and report of what has been done in the entire day, week, month related to his project he can also view the manpower details and how many people are working on his project, the status of the person who has been allotted for his project/product can be monitored.



Fig. [1] Software Development Life Cycle

As this software will provide help to the Software Development Industry, the scope of the project can also be extended to the other small scale and large scale industries.

OBJECTIVE

The paper primarily attempts to achieve the following:

- Establishing a standard for the better implementation of Project Management Software.
- Achieving maximum utilisation and user satisfaction in terms of accessibility and functionality of Project Management Services.

CURRENT MODELS AND NEED FOR IMPROVEMENT

Software systems tend to be complex and especially with the rise of service-oriented architecture, link multiple traditional systems potentially supplied by different software vendors. To manage this level of complexity, a number of SDLC models or methodologies have been used in the past, such as:

- Waterfall Model
- Spiral Model
- Rapid prototyping
- Incremental
- Synchronize and stabilize
- Agile software development

Though most of these models have worked to the fullest of their capabilities, the problem remains in many of the following issues:

- Satisfying compatibility constraints of various services from various vendors
- Adequate access of resources and managing them accordingly
- Managing the responsiveness and updating of changes required
- Failure of co-ordination between the client and service provider to fulfill what was required
- Deviation from the expected product
- If the initial phases don't turn out as expected, it is highly inconvenient to turn back and re-establish the analysis and reuse the resources, hence causing a great deal of loss

- Unexpected long term and unprecedented results may be faced if a non verified algorithm is used for resource utilization

Keeping in mind the above drawbacks of vendor based SDLC system, the Cloud based service which incorporates the phases of SDLC would offer the following:

- Step by Step requirement analysis in the initial phase through the Cloud Service and saving of these specifications for further reference
- Design according to the Requirement Analysis and customer specifications stored
- Implementation in the customer based environment through cloud service
- Testing to detect possible un-optimized resource usage and deviation from sole purpose
- Updating the service to appease the dynamic requirements of the customer

MODEL APPROACH

Cloud computing is a paradigm shift in the computing world and brings more scalable, flexible and yet robust environment for online management. It provides almost instant access to the software and development environments, by providing multi-tenancy of the virtualized servers and other IT infrastructure. The following steps would be required to implement our Cloud service for effective Project Management:

Cloud Service

The Cloud Service would provide many advantages over the conventional Software Architecture:

- **Robustness:** Due to the data being centralized for the Cloud, resources are accessible to each end user provided they are authorized (hence providing a facility to make resources as secure as orthodox systems).
- **Platform Independent Access:** No constraints as to where one would be authorized to access the resources and progress of phases from. Could be authorized access from Mobile, PC, or other remote devices.
- **Scalability:** The scope of the project management can be virtually increased depending on where and what features are to be added to the software service on the cloud for a project.
- **Frequent Maintenance:** It is considerably easy to maintain and update a service as a software since the software doesn't need to be installed on the end where updates are being done from or from the end where maintenance strategy is being implemented.
- **Application Programming Interface:** Cloud Computing allows us to use API due to which we

increase reusability of component which helps in rapid development of software.

As Wu, Garg, and Buyya^[4] mentioned in their paper about the scope of SaaS (Software as a Service) and its possibilities to revolutionize the way end users make use of softwares provided as a service over the Internet. The model strictly follows a 'pay-as-you-go' concept which implies that one need not worry about the trivial issues related with the mundane softwares that are used commonly for project management phases, such as copyright violation and distribution or even piracy of the software. The SaaS would hence minimize cost requirements and provide end users with the service only when it is required; adding the functionality of not paying for what is not required; which adds favourability for the customers/end users.

Cloud based PMS's another important feature is, it implements agile software development model; which is of substantial importance since selection of software development model is also one of the main factors for software failure. Traditional water fall model doesn't give you the success rate of agile software development model.

In conventional project management systems, if one is not satisfied with an initial level phase of the SDLC cycle, the further stages which are already under implementation would have to be managed or even discarded accordingly. This increases the time and cost load for the client and this is where our SaaS services proves out to provide a better solution as discussed later in the paper.

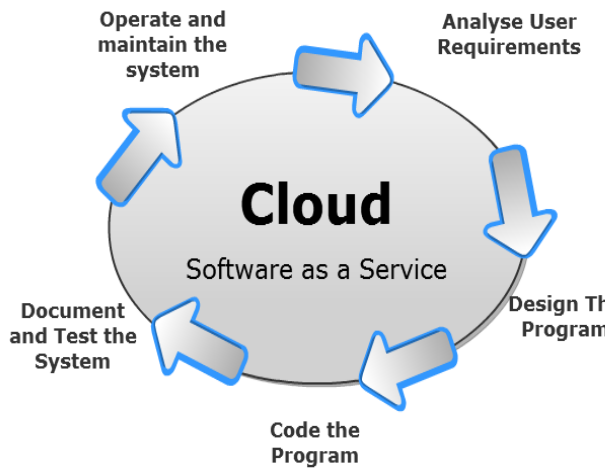


Fig. [2] Our proposed plan to incorporate the stages of SDLC through Software as a Service

Requirement Analysis and Development

The initial phase after the customer has registered for the SaaS on the respective Cloud would consist of the following steps:

[1] *Customer specifies Requirements:* This is where the end user/customer would specify the requirements, resources available, constraints to be kept in mind and what is to be done. These would be specified in the respective option of the Service Interface where the end

user would have to specifically list the above mentioned factors.

[2] *Service decision:* Once the requirements have been clarified by the customer, the Service as a Software would analyse whether the given constraints (such as deadline) and resources would be feasibly to satisfy the customer and if it is possible to do so by any optimization.

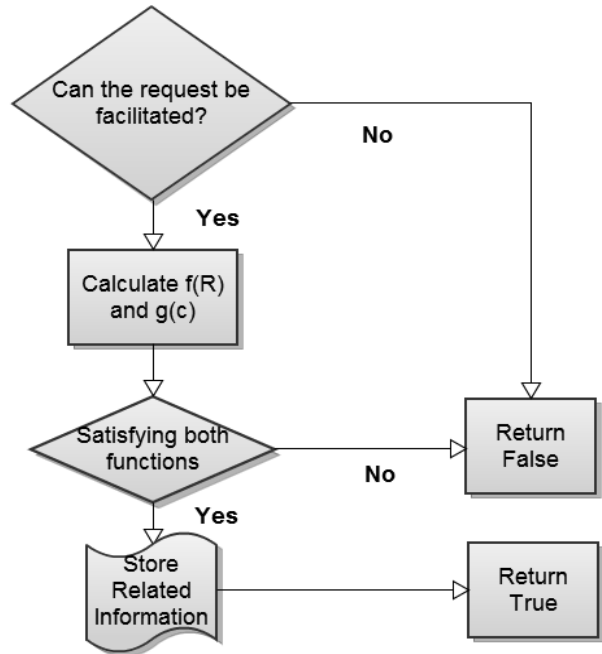


Fig. [3] Flowchart showing whether the job can be admitted keeping in mind the deadline or not

If the Service reckons that:

$$R \geq f(R)$$

(R – resources, c – constraints, f () – function to perform SDLC with the respective 'R', g () – function to be performed in order to check constraint boundaries pertaining to 'c')

And if g(c) is being satisfied, the Service would inform the user straight forward about the sustainability of the service for the given requirements and project. If no, then it would also inform the user accordingly and hence disable the further stages of SDLC for the given project, and discard it from the system if the user cannot accommodate the resources and specifications for the service to be made possible.

[3] *Utilization & Development:* The service would then make use of the SaaS layer most efficiently using scheduling algorithms based on which projects phase is to be served before (since a practical model of the system would be handling hundreds if not thousands of projects). This involves using the requirements accumulated in the first phase which in orthodox model is done by a third party. But since the main objective of this model is to minimize the tediousness, the requirements and resources would simply be passed on to the development and

utilization stage with the help of the SaaS and IaaS as we would see.

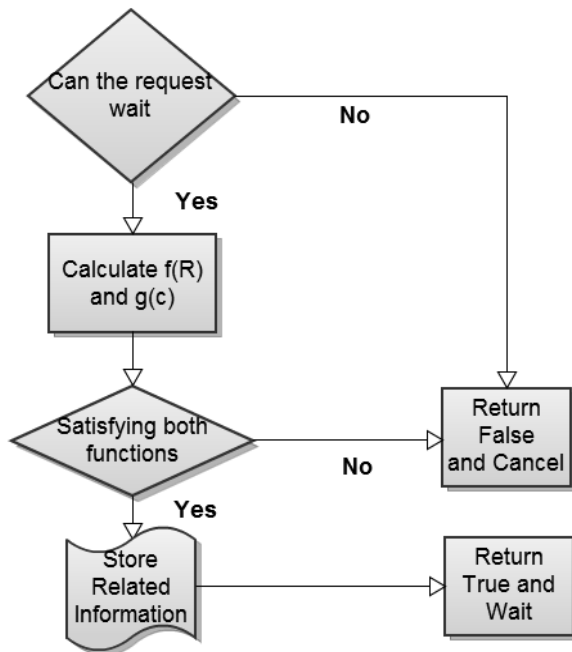


Fig. [4]. Flowchart showing the algorithm for WAIT state or CANCEL state

The waiting algorithm as depicted in the figure serves the purpose of checking if one request would cause the delay in a much more prioritized request further in time to be accommodated. Hence using the simple if else nested clauses it determines in the long run if a problem/loss would be encountered as to serving the respective requests at a particular time.

Basically a SaaS borrows services and resources from IaaS (Infrastructure as a Service) providers, and in turn leases those services to the users. The whole advantage of doing so is that SaaS maximizes resource utilization instead of giving the user total control of the services in terms of resources. This also results in increasing the Customer Satisfaction Level (CSL) and further considerations such as QoS (Quality of Service). However, it is important to have two layers of SaaS which would both function extremely differently in polarity but with co-ordination to assure a justified use of resources. These layers will be discussed further in this section.

The SaaS would then borrow the resources of IaaS (Infrastructure as a Service) and make it possible for the development and design of the specified project. An IaaS is used to dispatch the core services to the SaaS for making it possible to implement the functioning of a Virtual Machine. In this process, the resources would be carefully and meticulously observed and used up. It is to be noted that the platform layer of SaaS creates instances using the VM images.

To put it in layman terms, the user only deals with SaaS and the SaaS serves the user by borrowing services from

ends known as Infrastructure as a Service (IaaS) which lends the respective Services for a specific task or job.

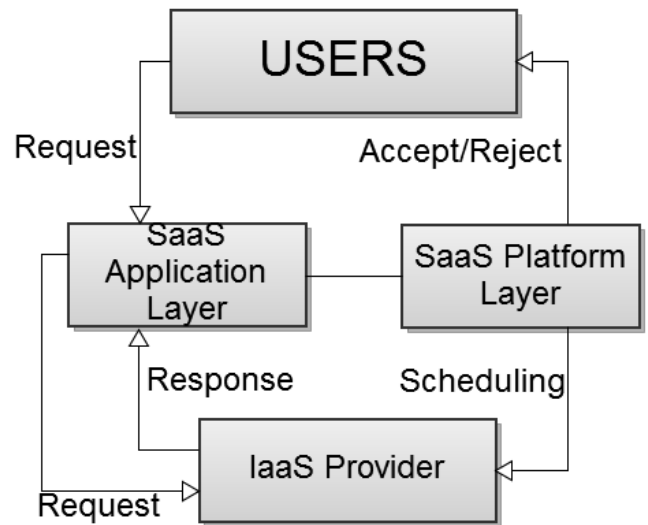


Fig. [5] Representation of the internal working through the two Services - IaaS and SaaS

The SaaS contains two vital layers:

1) The Platform Layer which is responsible for the Admission Control depending on how many projects are already admitted, and the scheduling process for which we consider many options like Costs for waiting, Investment Return.

2) The Application Layer which is required to assemble the service from IaaS and integrate the resources with it to perform the job which otherwise would have been done by a third party system in an orthodox system.

~~Maintenance and Updating~~

The most important phase to ensure and endure the long lasting accuracy and compatibility of a software would be its maintenance and report generation. Once the project development has been done, it is to be tested at the user end and reported on a diurnal basis at the Service end. The Service would then use appropriate facilitation of the IaaS to generate reports of the project performance on a regular basis.

This in turn would allow us to manage the report generation phase without co-ordinating with a third party and with only contact needed with the end user/customer. The company may not even have to contact the respective officials related to the Cloud Service but just Log-In and manage their maintenance routine without any prior assistance of the Service providers.

Reports would hence be produced and saved for archive resources of each project, just so if a user needs to observe a pattern of performance of the system, the Services

would be able to procure him with the requisite reports anywhere, anytime on the Cloud.

CONCLUSION

We kept into consideration the time and resource limitations of the customer and yet managed to prove an efficient usage of a cloud service based project management system without any third party direct interference. We made sure pre-admittance of the project whether the project would be accommodated by the Service on the Cloud given the constraints and resources by the user and inform them right away, to not keep them disembarked on the wrong path and provide an earnest service. We also implemented scheduling of processes to keep the ambiguity of when a project would be allowed to be serviced by the intended service.

In the end, a powerful and robust system which would facilitate the SDLC requirements of an end user is accomplished to be used without unnecessary interference from any other vendor.

Moreover, this model can be implemented not only for small term developer level projects but also higher level project management for which the number of resources to be utilized are a long term and non-ephemeral function of usage and maintenance.

FUTURE DEVELOPMENT

The robustness and scalability of such a Cloud Service based platform for Project Management Software should make the process of managing, development and maintenance of projects in the future a more automated yet still a highly cohesive for project management at all levels. Hence, a company/user would be able to avail the same services that a company of higher level avails at the same rate and level of quality regardless of whether the project is for a personal purpose or a highly professional purpose on a multi-national scale. The possibilities of augmenting other utilities to such a system would always be possible along with multipurpose functionalities which may rise with the further changes and advancements in Cloud Computing in the future.

ACKNOWLEDGMENT

We would like to express our sincere gratitude towards Prof. Ahmad Raza Khan for guiding us throughout this paper research and providing us eclectically the resources needed for us to make this paper possible.

REFERENCES

- [1] Paulo Ditarso, Francisco Brasileiro, Ricardo Araujo Santos, David Candeia, Raquel Lopes, Marcus Carvalho, Renato Miceli, Nazareno Andrade, Miranda Mowrbay, "Business-driven short-term management of a hybrid IT infrastructure", *Journal of parallel computing* [2012]
- [2] Gerald Ganold, Sudhakaran Mudiam, Timothy Lindquist, "Automated support for service-based software development and integration." *Journal of Systems and Software* 74 [2005]
- [3] Valentina Casola, Antonio Cuomo, Massimiliano Rak, Umerto Villano, "Future Generation Computer Systems", *Journal of Computer and System Sciences* [2008].
- [4] Linlin Wu, Saurabh K Garg, Rajkumar Buyya, "SLA based admission control over SaaS in Cloud computing environments" *Journal of Computer and System Sciences* [2011]
- [5] Hongtao Du, Zhanhuai Li, "Online backup System for Cloud Computing Storage", *SciVerse Science Direct – Energy Procedia* [2011]
- [6] Salekul Islam, Jean-Charles Gregoire, "A Flexible Cloud Model and its application for Multimedia", *SciVerse Science Direct – Future Generation Computer Systems* 28 [2012]
- [7] Javier Garcia, Antonio Amescua, Maria-Isabel Sanchez, Leonardo Barmon, "Design guidelines for software process knowledge repository development", *Information and software Technology* 53 [2011]
- [8] Johan Tordsson, Ruben S. Montero, Rafael Moreno-Vozmediano, Ignacio M. Llorente, "Cloud brokering mechanisms for optimized placement of virtual machines across multiple providers", *Future Generation Computer Systems* 28 [2012]
- [9] Andrew Joint, Edwin Baker, "Knowing the past to understand the present - Issues in contracting for Cloud based Services", *Computer Law and Security Review* 27 [2011]
- [10] Dana Petcu, Georgiana Macariu, Silviu Panica, Ciprian Craciun, "Portable Cloud Applications – from theory to practice", *Future Generation Computer Systems* 2012
- [11] Xie Lin Lin, Le Yun, Li Yong-Kui, Ning Yan, "A Study on the Management System Design for Large Complexity Projects", 4th International Conference on Computer science and Convergence Information Technology [2009]

Research on Forecasting Call Center Traffic through PCA and BP Artificial Neural Network

Tao Liu¹, Lieli Liu²

School of Economics and Management
Beijing University of Aeronautics and Astronautics
Beijing 100191, China
e-mail: liutaoterry@163.com, liulieli@yahoo.com.cn

Abstract—Accurately forecasting future call volumes is critical for scheduling of a call center. This thesis develops a PCA-BP model to forecast future call volumes of half-hour periods. The approach adopted firstly uses principle component analysis to eliminate the intraday correlations between the call volumes of 48 consecutive half-hour periods and to simplify the structure of BP neural network by dimension reduction. The processed data are then input into BP network for training. We use the trained network to forecast future call volumes and apply a competing model to the same data. It turns out that the new model performs better and can be adapted in call center traffic forecasting. To the best of our knowledge, the forecasting method we built has not been used in this area hitherto and it deserves trial application accordingly.

Keywords—call center; forecasting; principal component analysis; BP neural network

I. INTRODUCTION

Nowadays, lots huge call centers consist of thousands of agents. In the long run, the scales of call centers are bound to expand along with the explosion of modern business between enterprises and their customers. Under this circumstance, traditional methods like scheduling merely based experience will no longer meet the need of realistic production and operation, as traditional methods not only pile the workload with poor efficiency but also gap the deviation from reality. Therefore, many researchers have recently presented different alternative methods and models in different research areas such as forecasting, queuing, and scheduling.

According to Aksin et al. [1], call centers are labor-intensive operations, with the cost of agents generally comprising 60-80% of the overall operating budget. Thus, the most effective management strategy requires call center managers to schedule appropriate amount of agents dealing with phone calls. To do this, the first and most critical step is to accurately forecast future call volumes before scheduling. These years, linear time series models are mostly introduced to forecast future call volumes. Early work on forecasting call volumes usually applies standard time series methods, such as autoregressive integrated moving average (ARIMA) models, which are usually used in forecasting call center traffic (Andrews and Cunningham [2]; Bianchi et al. [3]). Recently, alternative time series models are introduced abundantly. Taylor [4] introduced some of these models and compared their

forecasting accuracy. These models regard historical call center arrivals as linear time series data, smoothing the time series data with weighted average of past observations, or fitting the linear mathematical models by performing regression and statistical test to estimate parameters. However, as large amount of time series are nonlinear time series, the forecasting performance by applying linear time series models are usually not satisfying.

In recent years, along with the rapid development of artificial intelligence, many advanced algorithms have been invented to solve all kinds of problems concerned such as forecasting and classification. Artificial neural network (ANN) is an artificial intelligence algorithm which mimics the way signal propagating through brain nerve cells. Back propagation (BP) neural network is the most commonly used one, especially in the area of forecasting. If the structure is constructed appropriately, BP neural network can fit almost every complex nonlinear function with its extraordinary approximation and mapping capabilities. However, BP neural network is rather sensitive to the structure. The determination of the amount of variables for the input layer makes great influence on the performance of the algorithm. Excessive variables make the network structure complicated, resulting in the aggravation of training burden and the shrink of convergence speed in learning. Meanwhile, the lack of correlations between some selected input and output variables is liable to drop into local minima, which will surely lower the performance of forecasting. As there are many lying factors may influence future call volumes, only the right combination for input can output precise forecasting results.

This paper firstly proposes a principal component analysis (PCA) technique to extract the main impact factors that influence call volumes, and input those principal components into BP neural network for running the simulation and forecasting future call volumes. The rest of this paper is structured as follows. Section 2 and §3 respectively introduces the algorithm of PCA and that of BP neural network; §4 delivers the simulation run in MATLAB and performance of forecasting results. The paper conclusion is made in §5.

II. PRINCIPAL COMPONENT ANALYSIS

Principal component analysis (PCA) is a multivariate statistical analysis method. PCA can extract the principle

components from multiple-related variables by calculating the covariance matrix (Jolliffe [5]). These principle components (PCs) are independent from each other. They conclude the main information of the original variables, while the amount of PCs is much fewer than the original variables. Thus, PCA is an effective tool in dimension reduction. The principles of PCA are as follows:

Suppose that n -dimension matrix $X = (x_1, x_2, \dots, x_n)$ is comprised of n variables. The linear combinations are:

$$\begin{cases} y_1 = c_{11}x_1 + c_{12}x_2 + \dots + c_{1n}x_n \\ y_2 = c_{21}x_1 + c_{22}x_2 + \dots + c_{2n}x_n \\ \vdots \\ y_n = c_{n1}x_1 + c_{n2}x_2 + \dots + c_{nn}x_n \end{cases} \quad (1)$$

where c_{ij} are decided by the following qualifications:

1) $c_{i1}^2 + c_{i2}^2 + \dots + c_{in}^2 = 1$ ($i = 1, 2, \dots, n$);

2) For $i \neq j$, y_i is independent to y_j ($i, j = 1, 2, \dots, n$);

3) y_1 is the linear combination of x_1, x_2, \dots, x_n according to (1) with the maximum variation, y_2 is the linear combination of x_1, x_2, \dots, x_n with the second maximum variation, and is not related to y_1 . In the same way, y_n is the linear combination of x_1, x_2, \dots, x_n with the minimum variation, and is not related to y_1, y_2, \dots, y_{n-1} .

Then, y_1, y_2, \dots, y_n are the n PCs that are extracted from the original variables. The quantity $\lambda_i / \sum_{k=1}^n \lambda_k$ ($i = 1, 2, \dots, n$) measures the relative importance of the i th PC. The cumulative importance of the first m PCs is defined as $\sum_{i=1}^m \lambda_i / \sum_{k=1}^n \lambda_k$. If the cumulative importance is beyond 80%~90% (90% in this paper), dimension reduction is performed by replacing the original variables with these first m PCs in analysis.

The main procedures to carry out PCA are shown below:

1) To standardize the original data. In order to eliminate the impact of the large numerical difference and the difference in dimension of the original variables, the original variables have to be standardized.

$$y_{ij} = \frac{x_{ij} - \bar{x}_j}{s_j} \in Y \quad (2)$$

where \bar{x}_j is the mean value of j th column, and s_j denotes the variance of j th column, and $x_{ij} \in X$, $i = 1, 2, \dots, m$, $j = 1, 2, \dots, n$.

2) To establish the correlation coefficient matrix R .

$$R = \frac{1}{n-1} Y'Y \quad (3)$$

3) To calculate the characteristic values λ_i ($i = 1, 2, \dots, p$) from the characteristic equation $|\lambda_i - R| = 0$, then get the corresponding eigenvectors e_i ($i = 1, 2, \dots, p$).

4) To determine the first m PCs which make the cumulative importance $\sum_{i=1}^m \lambda_i / \sum_{k=1}^n \lambda_k$ beyond 90%.

5) To calculate the principal components loads ∂_i and principal components scores F_i .

$$\partial_i = \sum_{i=1}^m \sqrt{\lambda_i} e_i \quad (4)$$

where e_i are the eigenvectors of the characteristic values λ_i ($i = 1, 2, \dots, m$).

$$F_i = y_i \partial_i \quad (5)$$

where y_i is the standardized value of x_i .

III. BP ARTIFICIAL NEURAL NETWORK

Artificial neural network (ANN) is a kind of intelligent information processing technology which simulates the information processing and storage mechanism of human brains. In essence, it realizes a mapping from input to output. The output is determined by input samples, transfer functions, and connection weights between neurons. The mapping from input to output is fulfilled by modifying the connection weights through training and learning process.

Back Propagation (BP) neural network is one of the most widely used and mature artificial neural networks. It is a kind of feed-forward neural networks with supervised learning mechanism. According to the Kolmogorov's mapping neural network existence theorem (R. Hecht-Nielsen [6]), it proves the existence of a three-layered network, whose hidden unit output function is a nonlinear function, and the functions of input and output unit are linear functions. The relationship between input and output of this network can approximate any nonlinear function. Thus, the BP neural network model this paper built is a three-layered structured one, which is comprised of an input layer, a hidden layer, and an output layer.

Our BP model is realized by the following steps:

1) To divide samples. The network inputs are the first m dimensions of principal components scores matrix from PCA in §2. The corresponding outputs are the call volumes of specific unit time intervals (30 minutes). These samples should be divided into training set and testing set.

2) To standardize both data sets into $[-1,1]$.

$$y_{ij} = 2 \frac{x_{ij} - X_{\min}}{X_{\max} - X_{\min}} - 1 \quad (6)$$

3) To initialize w_{ij} , w_{jk} , θ_j , and θ_k with random values in $(-1,1)$. Where w_{ij} are the connection weights between each units from the input layer and the hidden layer, w_{jk} are the connection weights between each units from the hidden layer and the output layer, θ_j are the output threshold of each units from the hidden layer, and θ_k are the output threshold of each units from the output layer.

4) To find a group of variables (X_t, T_t) in the training set randomly, where X_t are the input variables and T_t are the real call volumes as output variables.

5) To calculate the output values for each units from the hidden layer.

$$Y_j^2 = f\left(\sum_{i=1}^{n_1} W_{ij} Y_i^1 - \theta_j\right) = f\left(\sum_{i=1}^{n_1} W_{ij} X_{ti} - \theta_j\right) \quad (7)$$

$$i = (1, 2, \dots, n_1) \quad j = (1, 2, \dots, n_2)$$

6) To calculate the output values for each units from the output layer.

$$Y_k^3 = f\left(\sum_{j=1}^{n_2} W_{jk} Y_j^2 - \theta_k\right) \quad k = (1, 2, \dots, m) \quad (8)$$

7) To calculate the errors between Y_k^3 and T_t . If the errors are in the permitted scope, the training period is over. Otherwise the connection weights w_{ij} and w_{jk} have to be modified.

8) To test the trained network with the testing data set. If the errors between inputs and outputs are acceptable, then the model can be well used to forecast future call volumes.

IV. EMPIRICAL STUDY

In this section, actual data are used to show the implementation of the model we proposed. Section 4.1 introduces the data source and a preliminary data analysis. The process the PCA-BP model simulated in MATLAB with actual data are described in §4.2, and is followed by §4.3 which focuses on the forecasting performance as the research results.

A. Data

The present data were gathered at the call center of a major telecommunication company in China. They were collected over 366 days from April 1, 2011 to Mar 31, 2012. Call volumes of each day are aggregated in consecutive time intervals of 30 minutes each. There are 48 intervals in a day, and a total of 17,568 observations in the data set. As the call arrival pattern is quite different between weekdays and weekends or holidays, the experiment was focused solely on weekdays. Thus a total of 116 weekends or holidays were removed from the data sets, with 250 remaining days left and 12,000 observations kept in the data set.

B. Forecasting Process

Analyzed by the paper, all weekdays have a similar intraday profile, namely, call arrival patterns for each weekday are almost the same and there is a periodicity of 48 consecutive time intervals of 30 minutes. Such intraday call arrival patterns are quite common in call centers; e.g., see Gans et al. [7]. In addition, correlations are relatively strong between successive weekdays, and are slightly smaller with longer lags. This phenomenon is also shown in Brown et al. [8]. Thus, the input of the paper-built model is a $n \times m$ matrix that records the call volumes for n groups of consecutive m time intervals of 30 minutes, and the corresponding output for each group is the call volume of the successive time period. In this case, $n = 12000 - 48 = 11952$, and $m = 48$.

According to Tanir and Booth [9] and our own analysis as well, there are strong correlations between call volumes of different time periods in a day, especially between successive periods. Therefore, in order to eliminate the intraday correlations between the call volumes of 48 time periods for each group and simplify the structure of input layer for BP neural network, PCA is conducted with MATLAB. We use *princomp()* to extract 9 PCs from the 48-dimension input matrix. Shen and Huang [10] used a similar method of dimension reduction. It is called the singular value decomposition (SVD) approach closely related to PCA when PCs are calculated from the covariance matrix.

The 9-dimension score matrix is then regarded as the input of BP neural network. Thus, there are 9 units in input layer. The number of hidden units is determined as 19 by empirical equation $N = 2M + 1$. The output layer has one unit containing the result of each time period. The data set is divided into training set and testing set. *newff()* is used from MATLAB to train the network with data from training set. According to the Kolmogorov theorem discussed in §3, *tansig* is chosen, which is a hyperbolic tangent sigmoid transfer function to be the activating function between input layer and hidden layer, and *purelin*, which is a linear transfer function to be the activating function between hidden layer and output layer. A developed training function *trainlm* is adapted, which is based on the Levenberg-Marquardt algorithm, offering the fastest convergence speed.

The testing data set is used to test the trained BP model. 400 pairs of Actual call volumes and predicted values are compared in Fig. 1.

C. Forecasting Performance

Mean absolute percentage error (MAPE) is applied to evaluate the overall forecasting performance.

$$MAPE(n) = \frac{100\%}{n} \sum_{i=1}^n \left| \frac{\hat{x}_i - x_i}{x_i} \right| \quad (9)$$

where x_i denotes the i th actual call volume, and \hat{x}_i is a forecast of x_i . For the errors between predicted data and real data, the smaller, the better.

The widely-used AR(1) model is also employed to the data collected. The comparison of AR(1) model and the paper-built PCA-BP model are shown in TABLE I.

Table I compares summary statistics of the MAPE of the forecasts from the AR(1) model and the paper-built PCA-BP model. It can be seen that the PCA-BP model is comparably better in forecasting performance.

V. CONCLUSIONS

Appropriate scheduling can elevate the efficiency and reduce the cost of a call center. The first and most critical step is to accurately forecasting future call volumes before scheduling. BP neural network performs better than linear time series models in nonlinear time series forecasting. However, BP neural network is rather sensitive to the structure. This problem was solved by adapting principle component analysis to eliminate the intraday correlations between input call volumes and therefore simplify the structure of BP neural network by dimension reduction.

This paper focuses on the forecast of weekday call center traffic. Forecasting call volumes of weekends or holidays and some emergency situations are not considered in this paper. However, our model may not perform well in those scenarios because of the insufficiency of historical data.

ACKNOWLEDGEMENT

The authors want to extend grateful thanks to Xiaoming Li, Min Huang, and Liangbao Zhang from China Satellite Communications Corporation Ltd., whose assistance in the phase of project implementation has greatly improved the scope and presentation of this paper.

TABLE I. Summary Statistics (Mean, Median, Lower Quartile Q1, Upper Quartile Q3) of MAPE

Model	MAPE (%)			
	Q1	Median	Mean	Q3
PCA-BP	2.3	4.8	5.6	8.4
AR(1)	9.5	13.9	18.8	23.8

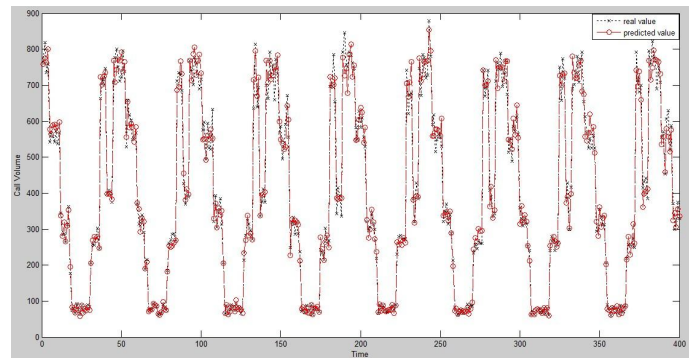


Figure 1. Contrast plot of actual call volumes and predicted values

REFERENCES

- [1] Z. Aksin, M. Armony, and V. Mehrotra, "The modern call center: A multi-disciplinary perspective on operations management research," *Production and Operations Management*, vol. 16, pp. 665-688, 2007.
- [2] B. H. Andrews and S. M. Cunningham, "L.L. Bean improves call-center forecasting," *Interfaces*, vol. 25, pp. 1-13, 1995.
- [3] L. Bianchi, J. Jarrett, and R. C. Hanumara, "Forecasting incoming calls to telemarketing centers," *The Journal of Business Forecasting Methods and Systems*, vol. 12, pp. 3-12, 1993.
- [4] J. W. Taylor, "A comparison of univariate time series methods for forecasting intraday arrivals at a call center," *Management Science*, vol. 54, pp. 253-265, 2008.
- [5] I. T. Jolliffe, *Principal Component Analysis*, 2nd ed., New York: Springer-Verlag, 2002, pp. 10-27.
- [6] R. Hecht-Nielsen, "Kolmogorov's mapping neural network existence theorem," *Proc. 1st IEEE Int. Conf. on Neural Networks*, IEEE Press, 1987, vol. 3, pp. 11-14.
- [7] N. Gans, G. M. Koole, and A. Mandelbaum, "Telephone call centers: tutorial, review, and research prospects," *Manufacturing and Service Operations Management*, vol. 5, pp. 79-141, 2003.
- [8] L. D. Brown, N. Gans, A. Mandelbaum, A. Sakov, H. Shen, S. Zeltyn, and L. H. Zhao, "Statistical analysis of a telephone call center: a queuing-science perspective," *Journal of American Statistical Association*, vol. 100, pp. 36-50, 2005.
- [9] O. Tanir and R. J. Booth, "Call center simulation in Bell Canada," *Proc. 1999 Winter Simulation Conf. (WSC 99)*, IEEE Press, Dec. 1999, vol. 2, pp. 1640-1647, doi: 10.1109/WSC.1999.816904.
- [10] H. Shen and J. Z. Huang, "Interday forecasting and intraday updating of call center arrivals," *Manufacturing and Service Operations Management*, vol. 10, pp. 391-410, 2008.

SoR based Request Routing for Future CDN

Janaka Wijekoon*, Erwin Harahap**, Hiroaki Nishi***

Nishi Laboratory, Graduate School of Science and Technology, Keio University, Japan

{*janaka,**erwin2h,***west}@west.sd.keio.ac.jp

Abstract—Internet users are constantly demanding faster and higher quality services from their internet service providers. This results an increase in broadband services, and access to delivery of applications enriched in data become top priority. Therefore, for fast delivery of such applications, Content Delivery Networks (CDNs) have been introduced. The Internet has become a massive system consisting of enormous amounts of information. To maintain the rich information in the Internet and to achieve maximum benefit from networks, Service-oriented Routers (SoRs) have been introduced. A SoR has a high-throughput database and is able to analyze all transactions on its interfaces. In addition, SoRs can provide APIs for accessing stored contents in order to enrich services. CDNs generally use Request Routing (RR) methods to achieve low user latency by performing load balancing among servers. RR is accomplished by selecting the nearest server for a particular user. In this paper we implement a new RR infrastructure that is based on the content of packet streams. The proposed system is independent of DNS-based RR, which is the predominant RR method used by current CDN infrastructures. In our system, content-centric request re-routing with SoR is used. Experimental evaluation and comparison of the round trip time (RTT) result of our proposed system with DNS-based RR indicates that our system achieves 50-60% reduction in connection initiation time. The SoR based RR method is able to reroute packets without disturbing clients, which is not possible using existing RR methods. This feature will make future CDNs more effective and convenient.

Keywords – *Service-oriented Router, Content Delivery Network, Request Routing, Packet Stream.*

I. INTRODUCTION

In the past few decades, communication technologies have developed significantly and new technologies and methods for effective communication have been invented. In addition, people have become more interested in sharing knowledge and information for business and academic purposes. This combination of both information sharing and communication technologies has resulted in sophisticated environments for effective communication and information delivery. In this type of environment, researchers continue to research new technologies for the development of information sharing and communication methods. In terms of effective communication technologies, the Internet is the most significant invention of humankind.

The popularity and scope of the Internet has increased in the past couple of decades, as more and more people tend to share information via the Internet. The Internet since become a massive data repository for millions of people. These data are exchanged among people by Internet-backbone routers. A

network router is a device that connects several independent networks and forwards data from source to destination. Regular routers do not yet have the capability of providing content-based services, and this leads to an implicit limitation in networks and limits the benefit of carriers as well.

Data has become more valuable and useful as content service providers create many services using data. Routers have also become invaluable devices for interconnecting networks and relaying data on the Internet. Therefore cooperation among service providers and Internet carriers is important for future Internet.

As stated before, a router is invaluable device for interconnecting networks. Routers can process any kind of information included in a packet stream and can passively capture all content or data. In our laboratory, as a next-generation Internet, a new model of Internet infrastructure based on Service-oriented Routers (SoRs) was introduced [1-3]. A SoR can provide services to end users from the router itself using defined APIs. It provides many benefits because it enables passive data collection. This is in contrast to current end-to-end systems, which have to collect data actively. In active data collection, end hosts can get the needed data only by accessing other hosts, such as the Web crawlers of search engines. This is time consuming and the coverage of the data collection is limited. Frequent crawling for obtaining real-time status of the Internet sometimes causes network congestion. Passive data collection by SoRs enables real-time data acquisition showing the current status of the Internet [2].

As users are constantly demanding fast information delivery, data need to rapidly traverse the Internet. Eventually information has come to be regarded as content and the delivery process of this content has made considerable commission for Internet engineers. Users are demanding faster and higher quality services and data-rich application, such as streaming services, banking services and online payment services, from their services providers. This circumstance exhausts broadband services. Thus for faster, more effective data delivery over the Internet, Content Delivery Networks (CDN) have been introduced. As a result, the Internet is more optimized to achieve low user-perceived latency. In a typical CDN, content is replicated over a number of servers connected to the Internet in order to optimize content traffic. The servers are placed in different geographical regions. An overloaded server in a CDN, reduces the number of clients that can be served and drastically increases the client response time as well. Hence, the commonly used method of achieving the performance objectives of a CDN is to use a good Request Routing (RR) mechanism that can perform server load

balancing and localize the requests within a geographical region close to where those requests originate.[10-12]

Therefore, finding the nearest surrogate server for a particular CDN client is important in the minimization of client response delay. In other words, connection initiation time is highly dependent on the RR method used by the CDN provider. In this paper, we propose a model for effective and efficient RR that uses the features of SoR.

A SoR has special features that have not been implemented in existing networks. Thus an evaluation of networks composed of SoRs is essential. The main function of a SoR is to monitor packet streams and conduct a deep inspection of packets including payload as well as headers. In addition, a SoR should reroute packets according to their payloads. We introduced new SoR modules to NS-2[5] to simulate SoR-based RR architecture. Our overall simulation for SoR-based RR for CDN infrastructure and comparison of DNS-based RR; consisted of the following: (1) enhancing NS-2 simulation environment to achieve SoR features, (2) conducting a detailed inspection of User Datagram Protocol(UDP) packet streams and maintaining an indexing table in NS-2 SoR, (3) implementing content-centric packet rerouting according to the payload that carried by packets and (4) implementing the DNS-based RR method in the NS-2 simulation environment and comparing it with the SoR-based RR method.

The remainder of this paper is organized as follows. Section II discussed existing rerouting methods in brief. This is followed by discussion of SoR CDN infrastructure in section III. Section IV illustrates the novel modules introduced to the NS-2 simulator to simulate SoR CDN infrastructure, while section V details the simulation and presents the results. Section VI concludes this paper with a summary of important results and prospective future works.

II. EXISTING REQUEST ROUTING METHODS

A. DNS-based RR

Domain Name Service (DNS)-based RR is a widely used method in the current Internet. In addition, DNS is used by many CDN services as a directory service. In this process, the client initiates the communication while the browser initiates name lookup for the nearest surrogate server in the local DNS. If the local DNS server cache misses, the local DNS server sends the request to the root DNS server. The root DNS server then returns the address of the authoritative DNS server for the Web site. In the next Step, the authoritative DNS server returns the address of a surrogate server close to the client based on routing, load balancing, and Internet mapping mechanisms. Finally, the client retrieves the content from the nearest surrogate server [7].

The DNS based RR method has several limitations. Since it involves multiple levels of redirections, if the local DNS cache is missed, DNS lookup takes a long round trip time to find the nearest DNS server, irrespective of the client location. In addition, if DNS uses a short time-to-live to respond quickly to network changes, it increases the load in the DNS servers. In fact, DNS servers experience difficulties scaling to support

multiple content provider networks and larger numbers of content providers [7].

B. Application-layer based RR

In the current context, several application-based RR mechanisms have been introduced [8,9]. However these RR mechanisms work in combination with DNS-based RR mechanisms and, as a result, they inherit some disadvantages of DNS-based RR mechanisms. This type of mechanism can be classified into Uniform Resource Locator (URL)-based and MIME header (or site-specific)-based methods. Two types of URL-based RRs are 302-Redirection [9] and In-path Element. MIME-based RR uses MIME header elements such as cookies and language to make the RR decision [7].

Another interesting feature of application-based RR is the fact that content providers can modify the content. Several contents are built with the structure of the actual content data and the reference point of that content. Using the advantage of the content reference or location, content providers can modify the reference point to the nearest surrogate server. Consequently, clients receive data from the nearest surrogate server. This technique is also known as URL rewriting [7,9].

In addition, there are several other types of application-based RR mechanisms used by CDN services. Those are application layer anycasting and URL forwarding. Due to space limitations, we will not discuss them in this paper.

C. Transport-layer based-RR

Transport layer-based RR methods are used to achieve fine DNS-based RR. In addition, they are used to achieve the next level of RR after the first step is performed by the DNS-based RR. At this layer, RR uses IP addresses and port numbers to allocate sessions to more appropriate and nearer surrogate servers [7].

In addition to the above-mentioned three RR methods, DNS-based, Application-based and Transport-layer RR, there are other methods also available in the current CDN infrastructure namely Network Address Translation (NAT)-based and content layer-based RR[9]

D. Akamai way of rerouting

In the discussion of CDN RR mechanisms, we have to pay attention to the existing CDN providers and their ways of performing the RR. Akamai plays a vital role in the current CDN market [11]. Therefore, evaluation of our proposed method by comparing it with the Akamai way of RR strongly validates our evaluation. The Akamai RR method is based on DNS layer RR. Akamai maintains two levels of DNS servers to find the nearest surrogate to a particular client. In addition it uses the typical root DNS and the local architecture to achieve RR effectively from its point of view.

The Akamai low-level DNS servers update every 20s to maintain consistency with in the network [2]. However, as we mentioned in sub section II-A, this makes the network more congested and unstable. Yamaki et al. [6] of Nishi laboratory conducted a study of the DNS traffic in SINET Japan, which showed that 13.6% of the traffic at any particular time is DNS.

Thus, it can be said that DNS packets consume 13.6% of the bandwidth in a network. Therefore, minimizing the amount of DNS traffic in a network will result in more effective utilization of the network. In addition, Akamai's way of RR inherits all the disadvantages of the DNS-based RR method.

After studying the current CDN infrastructures and the RR methods available for CDN infrastructures, we devised a new method that is completely independent of the DNS-based RR method. Simulation, implementation and test results are discussed in the following sections.

III. SoR CDN INFRASTRUCTURE

Our study is focused on a novel model for the RR mechanism of CDN infrastructures that is independent of DNS-based RR. For effective, uninterrupted services, CDN infrastructures have to support faster ways of initiating connections and routing traffic according to the capabilities of the CDN servers for serving clients. In other words, server load balancing should be precise. The main problems faced by existing approaches are response time for connection initiation and critical faults such as server failure. Existing approaches also take a long time to operate because they depend on some higher-layer protocols.

We propose a new way to resolve above two issues by proposing a new CDN infrastructure based on SoRs. Since SoRs can passively collect packet streams, they can manipulate packet streams without depending on higher-layer protocols. A SoR is sufficiently able to reroute packets according to the payload carried by that packet. As a result, packets can be rerouted their respective destinations using the SoR itself. Consequently, the infrastructure is entirely independent of DNS and other kinds of redirections methods.

Our proposed CDN infrastructure can be used to evaluate new methods of CDN redirection and CDN load balancing algorithms in order to provide faster, uninterrupted services to users. We evaluated our proposed CDN infrastructure using two techniques: (1) content-based dynamic packet redirection in connection initiation and (2) dynamic content migration in emergencies such as server failure and link failure. This was done using the proposed SoR modules in NS-2 simulation.

IV. NS-2 SIMULATION EXTENSION FOR SoR

Francesco et al [4] conducted CDN simulation for server load balancing using NS-2. They used only the typical methods available in NS-2 to perform their simulations. However, those simulation extensions could not sufficiently simulate SoR features. As a result, we propose new essential NS-2 modules to simulate SoR features in NS-2 simulation platform.

Owing to the importance of RR mechanisms and their effect on the overall performance of CDN infrastructures, we developed extensions for NS-2 that facilitate easy evaluating of new algorithms for SoR-based CDN RR. Those extensions are proposed as new modules for NS-2 to support SoR and CDN simulation.

1) NS-2 SoR simulation extension

The SoR extension, depicted in figure 1, consists of several modifications and novel implementations for the NS-2 simulation environment. We implemented a module to generate Constant Bit Rate (CBR) traffic for CDN simulation. A novelty of this module is that, it takes user-defined data from the simulation code and attaches it to packets. As a result, we were able to achieve actual Content Centric Networking (CCN) in the proposed CDN architecture.

Figure 2 shows the implemented transport layer agent used in the NS-2 simulation environment to handle user data. We introduced a new packet structure that is capable of handling data and some tracing information to the NS-2. The transport layer agent is capable of taking application data from the application layer and introducing some special tracing variables: packet generated time, packet ID, cumulative agent packet ID, etc. It then creates an IP packet and transfers it to the data link and physical layers. It is also able to receive data from the data link layer and dispatch it to the relevant application while writing all necessary information to a trace file.

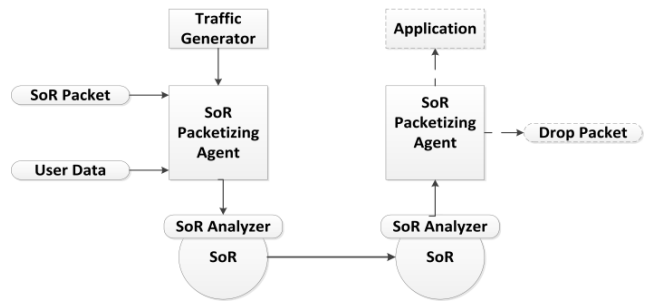


Figure 1. NS-2 SoR Extension

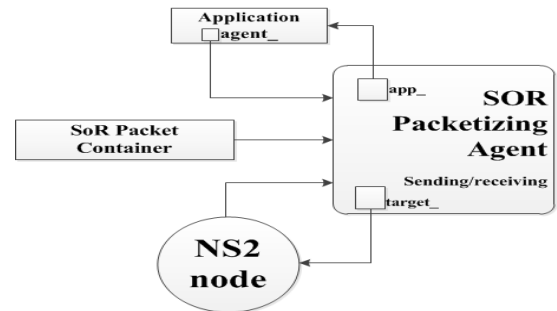


Figure 2. SoR Packetizing Agent

As illustrated in figure 3, we incorporated a new module into the NS-2 router node. The module is based on the principle virtual classifier of NS-2 [5] and is able to analyze each traffic stream entering the NS-2 router. It is also the basic building block of the SoR implementation of the NS-2 simulator. We added the ability to carry out Deep Packet Inspection (DPI) of packet stream data and dynamic rerouting according to the content carried by a packet. This feature opens up new path towards realizing the SoR-based CDN RR method.

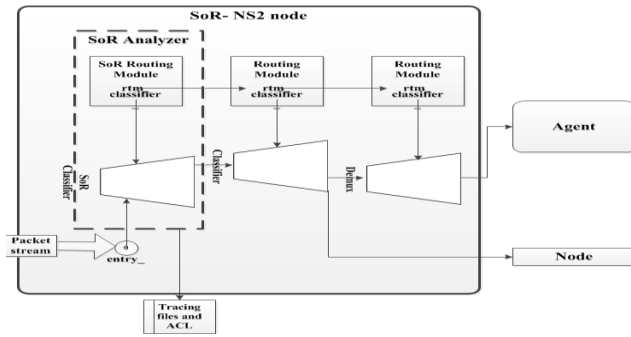


Figure 3. SoR Packet Stream Analyzer

V. SIMULATION TEST AND RESULT

A. Topology

To evaluate and compare the DNS and SoR-based RR methods, we used the topology depicted in Figure 4. The topology consists of one master server that is connected to router 2 and is capable of catering to the client the content “Site1” and “Site2”. In addition there are two surrogate servers connected to router 1 and router 4 supporting “Site1”, “Site2” and “Site2” respectively. DNS servers were also added to the simulation topology and arranged according to the Akamai way of DNS resolution for RR. For the DNS-based simulation all router nodes 0, 1, 2, 3 and 4 were typical NS-2 nodes, and for the SoR-based evaluation, all the routers were changed to SoRs.

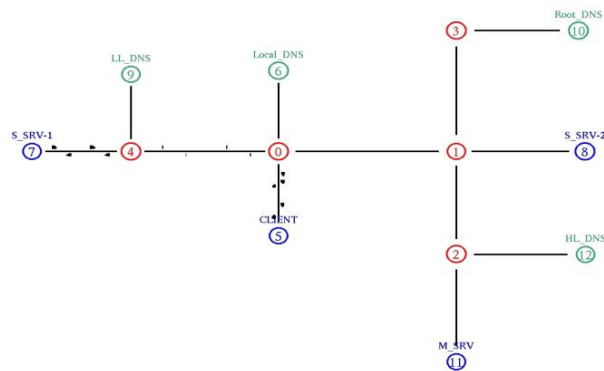


Figure 4. Simulation Topology

We made two assumptions: (1) all surrogate servers are up to date and (2) surrogate servers are placed in optimal positions. In addition, SoRs maintain a database that has a list of the nearest surrogate servers for some sets of data streams. In this experiment, we used CBR UDP traffic streams for simplifying session handling.

B. Simulation Scenarios

1) Content based redirection.

To evaluate the DNS and SoR-based RR scenarios, we simulated several traffic streams from a client. The client was sufficiently capable of sending several data traffic streams

defined in the Tool Command Language (TCL) simulation script. First, the client sent a request for “Site1”, and after a given interval of time, the client sent a request for “Site2.” Using those two scenarios, we evaluated the difference between SoR and DNS-based RR. We assumed that the local DNS cache is always missed. Therefore, at the beginning of the traffic streams, client should carry out the DNS lookup to determine the nearest surrogate server.

In the SoR-based RR architecture, once the client packet stream reaches its nearest SoR, the SoR analyzes the first packet of that traffic stream and searches the database for the nearest surrogate server for that traffic. It then changes the destination address of the packet stream to that of the nearest surrogate server. In addition, when we cached the properties of the first packet of a particular traffic stream, SoR changed the destination addresses of the remaining packets belonging to the traffic stream without database lookup. This scenario applied to all CDN data streams that passed through a SoR.

2) Dynamic connection migration

To simulate the connection interrupt, we developed a simulation scenario to send a message to the surrogate server for temporary interruption with possible locations of alternate surrogate servers. The surrogate server would then send a special message to the client about the possible surrogate servers. After receiving the message from the surrogate server, clients would change the destination address to that of the given surrogate server. In the DNS-based RR scenario, first, client should resolve the DNS. If the DNS cache hit is missed, connection re-initiation takes a long time to receive a response from the given surrogate server. In fact, in the SoR infrastructure, based on the second assumption, SoR will analyze the packet and directly redirect it to the corresponding surrogate server.

After running both simulations for 100s, we measured the time taken to receive a reply for each request. As shown in figure 5, at 0 s, the client initiated the request for “Site1” and the time taken to receive the reply for the first packet was above 0.25 s in the DNS-based redirection method. However, in the SoR-based redirection method it was below 0.15 s. This result clearly shows that our proposed method has great potential to speedily return replies to the user. At 10 s, the client initiated the request for “Site2” and as shown in figure 5 the DNS again took time to resolve the DNS for the nearest surrogate server, and the time taken to reach the first reply was significantly high. In contrast, in the SoR-based RR method, the time taken to reply is shorter than that in the DNS-based RR. This is because SoR forwards the packet for the nearest server by analyzing the content.

We then interrupted the connection between surrogate server 2 and the client, for “Site2” at 20 s, and then measured the time taken to redirect the packet stream to another surrogate server (surrogate server 1), for this connection interruption scenario. As local DNS cache hit missed, the client had to go through the DNS resolution process, and the time taken to receive a reply to the first request from new server was more than 0.25s. For the same scenario, however, the SoR-based redirection executed in a different way. The client sent the request to the new surrogate sarver directly. SoR identified the

first request and redirected it to the corresponding server, and the client was able to get a reply within 0.13s. In addition, we measured the packet loss in the connection re-initiation period and determined that 20-30 and 5-10 packets were lost in the DNS-based and SoR-based redirection respectively.

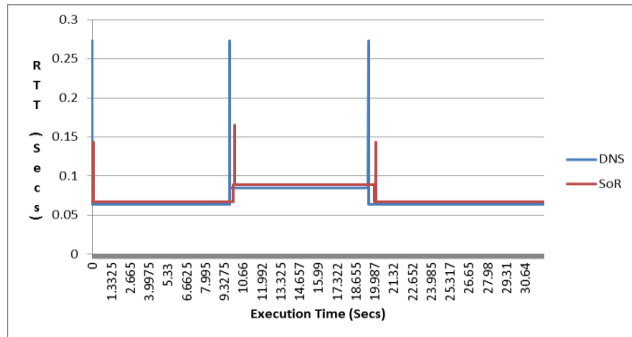


Figure 5. RTT Measurement at Client 1

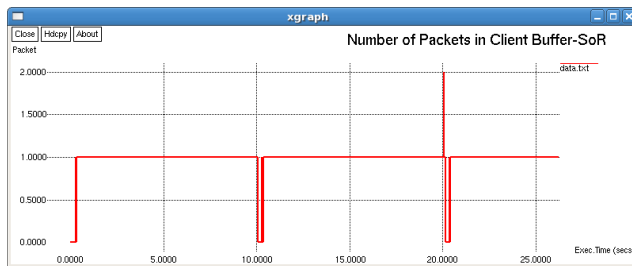


Figure 6. Client Buffer - DNS

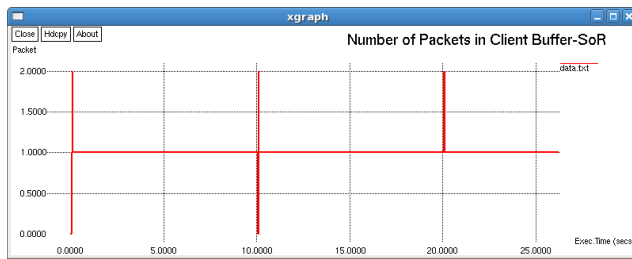


Figure 7. Client Buffer - SoR

As shown in figures 6 and 7, we measured the client buffer after each 0.0053s (same as the packet sending interval). The figures show that within the connection initiations and the redirections after 0, 10 and 20 s respectively, the client buffer was empty in the DNS-based RR scenario. The client connection was disrupted for a considerable length of time and the client had to wait a corresponding considerable length of time for connection initiation. In contrast, the same measurement for the SoR-based RR, showed the client buffer is empty only for a very short period of time, packet loss was low and client connection was almost uninterrupted. This proves that the hypothesis we made is correct and that we can achieve almost uninterrupted communication in SoR-based redirection.

As shown in figures 5, 6 and 7, the results we obtained from simulations indicate that SoR-based CDN RR architecture

has a considerable potential to make some revolutionary changes in CDN services.

VI. CONCLUSION

In this paper we introduced a novel CDN RR mechanism for CDN infrastructures. We identified some drawbacks in existing RR methods and showed how we are able to overcome these drawbacks using our proposed SoR-based approach. We also simulated the proposed SoR-based RR method and showed that it can significantly reduce the connection initiation time in a CDN. In addition, we showed that in a connection interruption situation, SoR can act fast enough to provide almost uninterrupted service to the client.

In the future we will test this method in an actual network with real network traffic. In addition, we intended to implement a model and a protocol to optimize the communication between SoRs and CDN servers, which addressed the server availability and the link state of the CDN architecture.

VII. ACKNOWLEDGEMENT

This work was partially supported by the National Institute of Information and Communications Technology (NICT) and a Grant-in-Aid for Scientific Research (C) (22500069), research was carried out at the Hiroaki Nishi Laboratory of Keio University, Japan.

VIII. REFERENCE

- [1] H.Nishi, "Service-oriented backbone router for future internet," Available: http://ec.europa.eu/information_society/activities/foi/research/eu-japan/prog/docs/day1stpm/architecture1-2/hnishi.pdf.
- [2] K. Inoue, D. Akashi, M. Koibuchi, H. Kawashima and H. Nishi "Semantic router using data stream to enrich services," 3rd International Conference on Future Internet Technologies (CFI08), Seoul, Korea, June - 2008.
- [3] H. Nishi, H. Kawashima, and M. Koibuchi. (2012) Information-based open innovation platform. Website [Online]. Available: <http://openinter.net/doku.php>
- [4] F. Cece, V. Formicola, F. Oliviero, and S. Romano, "An extended ns-2 for validation of load balancing algorithms in content delivery networks," in 3rd International ICST Conference on Simulation Tools and Techniques, March 2010.
- [5] K. Fall and K. Varadhan. (2011) The ns manual. [Online]. Available: http://www.isi.edu/nsnam/ns/doc/ns_doc.pdf
- [6] H.Yamaki, "Title of paper is known," unpublished.
- [7] M. Kabir, E. Manning and G. Shoja, "Request-routing trends and techniques in content distribution networks," in Proceedings of the ICCIT, Dhaka, December 2002.
- [8] Wang, Limin, Pai, Vivek, Peterson, Larry, "The effectiveness of request redirection on CDN robustness," Proceedings of the 5th symposium on Operating Systems Design and Implementation, June 2002
- [9] A. Barbir, B. Cain, F. Douglass, M. Green, M. Hofmann, R. Nair, D. Potter, O. Spatscheck, "Known CDN request-routing mechanisms draft-cain-cdn-known-request-routing-01," 2001.
- [10] X. He, S. Dawkins, Y. Zhang, "Routing request redirection for CDN interconnection draft-he-cdni-routing-request-redirection-", 2012.
- [11] Akamai. Akamai content delivery network. [Online]. Available: <http://www.akamai.com>.
- [12] Akamai Technologies, Fast Internet Content Delivery with Free Flow, April 2000.

Cognitive Radio for Adaptive Modulation and Coding

Sami H. O. Salih*, Abbas Mohammed*, Mamoun Suliman**

* Blekinge Institute of Technology, Karlskrona, Sweden

** Sudan University of Science and Technology, Khartoum, Sudan
szs@bth.se, amo@bth.se, mamounsuliman@yahoo.com

Abstract – Broadband Wireless Access (BWA) has become the best way to meet escalating business demand for rapid Internet connection and integrated "triple play" services. In addition, not only for topographic but also for technological limitations, alternative wireless solutions have been found. These systems are designed based on Cognitive Radio (CR) approaches, which can adjust its operation according to the environment and technical variations. This tracking feature allows the communication system to deliver the Best Ever, compare to Best Effort, services to the users. In this paper, an implementation of a cognitive engine for adaptive modulation and coding (AMC) is presented. This engine will track the radio channel variations in terms of SNR and be able to select a suitable modulation order among predefined Modulation and Coding Schemes (MCS) to maintain the specified BER by the user requirements.

Keywords: Cognitive Engine (CE); Adaptive Modulation and Coding (AMC); Software Defined Radio (SDR); Cognitive Radio (CR); Broadband Wireless Access (BWA)

I. INTRODUCTION

In traditional communication systems, the transmission is designed for the "worst case" channel scenario, thus coping with the channel variations due to the long term and short term fading or the Doppler effect and still delivering an error rate below the specified limit. Adaptive transmission schemes, instead, are designed to track the instantaneous channel quality and adapting channel throughput to the actual channel state. The tracking feature may relate to the signal strength (RSSI, IT or, SNR), the amount of traffic produced (Congestion, Erlang), the mobility profile of the Customer Premises Equipment (CPE), or a combination of two or more features [1, 2]. These techniques are taking the advantage of the time-varying nature of the wireless channel to vary the transmitted power level, symbol rate, coding scheme, constellation size, or any combination of these parameters, with the purpose to provide specific BER and hence improving the link average spectral efficiency measured by bits/s /Hz.

The concept of adaptive modulation was widely addressed, in [3, 4] the positive impact of using Adaptive Modulation and Coding (AMC) was demonstrated in terms of Quality of Service (QoS) metrics. In [5] a cognitive approach proposed using Fuzzy logic and in [10] neural-networks algorithm was used, however when considering the limited number of Modulation and Coding Schemes (MCS) profiles used Artificial Intelligent (AI) technique may inefficiently consume the system resources.

Furthermore, different simulation platforms were developed to consider specific parts of the Cognitive Radio

(CR) system, each of which simulates a part of the CR system, so it's not possible to aggregate different simulation platforms to investigate the performance of an entire CR system. Most of the developed test beds use Field Programmable Gate Array (FPGA) near radios (i.e. Antenna, Power Amplifier, and Frequency Band Converters) and general purpose devices to control the system [6, 7]. However, splitting the CR platform adds more complexity to the test bed since a high-speed connection is required between the two parts. None of the published papers, up to the best of the author's knowledge, address the issue of implementing a unified test bed simulating the behavior of the AMC function working for the access layer in BWA systems. Indeed, the Cognitive Engine (CE) function implemented here has two unique features; it's fully backward compatible with the Software Defined Radio (SDR) platform supporting any modulation order [1], and it's developed on modular approach, making it forward compatible with future MCSs.

This paper is structured as follows: In section II, an overview of the cognitive engine is discussed with relation to the previously implemented SDR function for AMC [1]. Design steps and the performance analysis of the CE design and its behavior on real world environment is demonstrated in sections III and VI, respectively. Finally, conclusions are drawn in section V.

II. ADAPTIVE COMMUNICATION SYSTEM DESIGN

A. Design Scope

For the time being, most BWA communication systems support variety of MCSs and allows for the schemes (also called profile) to change on a burst-by-burst basis per link, depending on channel conditions. Current systems contain separate hardware channel for each MCS. The more intelligent approach is to design a reconfigurable platform in hardware to support a wide range of channel specifications (for instance, WiMAX uses BPSK, QPSK, 16QAM, and 64QAM, or even higher constellations size in future systems) based on SDR, and then design a CE to determine which profile (or scheme) to load and operate. The AMC module is responsible for such mechanism, as shown in Figure 1, a higher data rate can be achieved by 64QAM in WiMAX when the channel conditions is relatively good. On the other hand, when the channel quality degrades, applying higher constellation will lead to dramatically increased the BER. Thus, a lower modulation order must be used to deliver reasonable Quality of Service (QoS) to the users.

On the half-duplex systems, a feedback channel is required to allow the mobile terminal update the base station

with the downlink Channel Quality Indicator (CQI). For the uplink, the base station can estimate the channel quality based on the received signal quality. In contrast, full-duplex systems, which is the case here, where the transmitter and receiver circuits are presented on both communication ends, the receiver part can perform the downlink channel quality measurement and pass the CQI to the transmitter locally, so there is no need for channel quality feedback, hence saving channel resources.

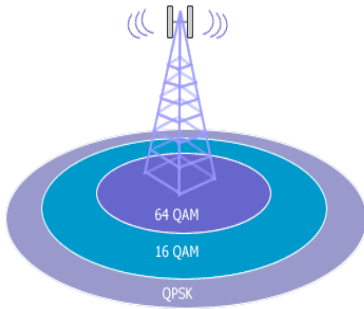


Figure 1. Adaptive Modulation and Coding used in WiMAX

B. SDR Platform

The BWA model is developed based on the IEEE 802.16e standard documents using SDR approach and it was successfully tested in [1] and the references thereon cited therein. A single function that can give different modulation order from BPSK to M-QAM ($M=2n$, where $n=2, 4, 6, \dots$) is implemented in Matlab. The function is called with the modulation order and the Signal-to-Noise Ratio (SNR) in dB as input, and then it plots the associated constellation and calculates the Bit Error Rate (BER) of the transmitter.

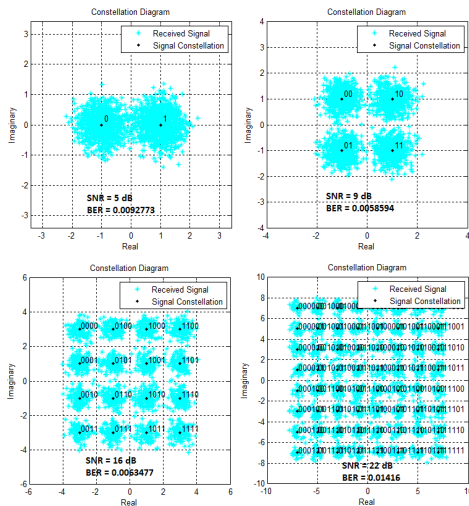


Figure 2. AMC Constellation Diagrams for BPSK, QPSK, 16QAM, and 64QAM

As Figure 2 represents when there is a good channel conditions in the communications link at a particular time higher modulation order can be used to improve the channel efficiency while maintaining an acceptable BER. Figure 3 shows the relation between SNR and the BER for various modulations order. From the graphs in Figure 3, in order to maintain a certain level of BER (transmission quality) for a given SNR (channel condition), a suitable modulation order have to be chosen to deliver the highest possible transmission quality.

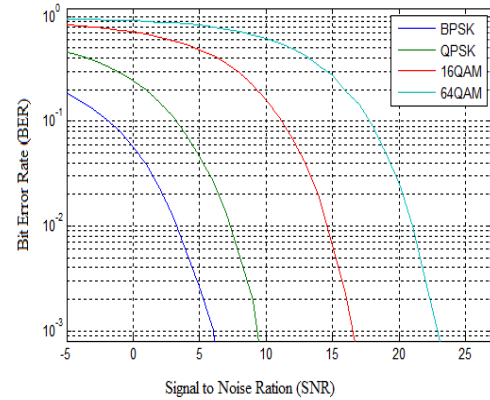


Figure 3. BER vs. SNR Lookup Table

Both functionality and performance of the AMC module can be compared with the hardware model in [2].

C. Cognitive Engine Functionality

The functionality of the CE is to continuously monitor the environment, by sensing selected cognitive features, refer to predefined policies, perform its logic to pick up the suitable configuration profile, and then automatically direct the SDR to load and execute the appropriate profile. Figure 4 shows the block diagram and the signal flow in the CE procedure.

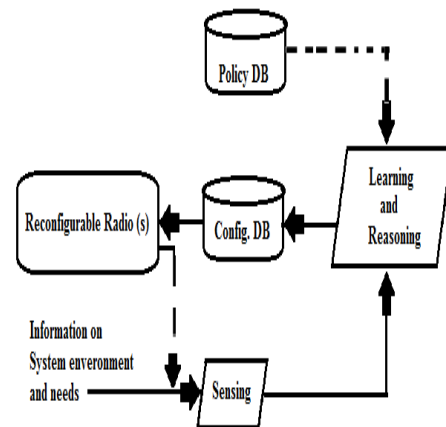


Figure 4. Cognitive Mechanism [3]

III. COGNITIVE ENGINE DESIGN

A. Design Environment

In this paper, Matlab functions are implemented to simulate the behavior of CE for the AMC element of the BWA systems. The implemented functions are fully compatible with a previous function implemented to simulate the SDR platform for AMC [1].

B. Cognitive Engine flowchart

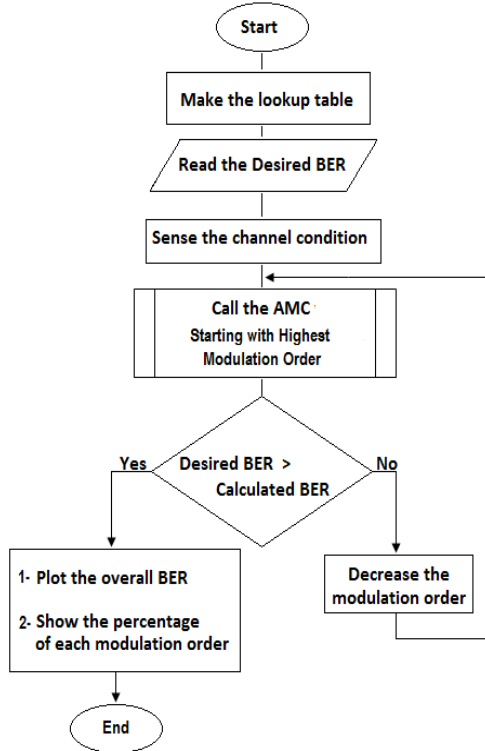


Figure 5. AMC Cognitive Engine Flowchart

This module is designed to work with the AMC function; together they perform the Cognitive AMC processes as shown in Figure 6.

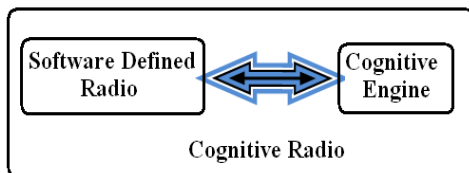


Figure 6. Cognitive Radio Module

Depending on application, the user (or the application layer protocol) has to provide the desired BER for the transmission; CE will perform its procedure to keep the BER below the targeted level. A lookup table is built representing the relation between SNR and BER as shown in Figure 3 to reflect the channel condition given in SNR to its vis-à-vis

BER. Hence, a logical comparison is made to select a suitable modulation order. By doing so, the maximum limit of BER is imposed by the user's desired BER as shown in Figure 7 (B, C, and D).

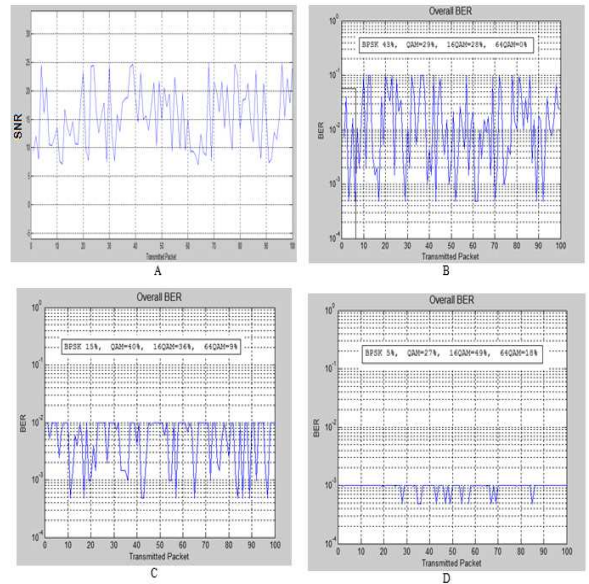


Figure 7. CE Functionality with Different User Desired BER

C. Performance Analysis

A random SNR is generated to examine the CE functionality as in Figure 7 A. In this scenario, the developed function is called three times with the user desired BER of 10^{-1} , 10^{-2} , and 10^{-3} as shown in Figure 7 B, C and D, respectively. The CE procedure has to decide which modulation order achieves the desired BER according to the channel condition as indicated by the SNR shown in Figure 7 (A), in order to demonstrate its functionality. CE function is successfully strictly complying with the maximum specified BER and adapt the transmission modulation order accordingly.

IV. SYSTEM IMPLEMENTATION

A. System Limitations

Adaptive modulation is more suitable for a two way communication system (Full Duplex) since the adaptive coefficients have to be synchronized in order to allow channel measurements and signaling to take place.

A typical communications system consist of three main parts: a transmitter that send the data (information), the receiver that process and retrieves the information, and the channel which is the physical media between the transmitter and the receiver. The CE part on the receiver circuit evaluates the received packets to estimate the CQI, and then adapt the SDR on the transmitter circuit according to a predefined logic. The unified test bed assumes that both SDR and CR parts are implemented on the same hardware

platform and at both communication ends as shown in Figure 8; this hypothesis gives the CE functions a direct accessibility to the all parameters within the transceiver which, with high probability, have the same values on the other end under the same channel conditions. Thus, no signaling channel is needed in parallel to the traffic channel, thus saving the radio spectrum channel resources. However, an internal signal flow is required in order to keep the transceiver harmonized (i.e. use the same modulation and demodulation order).

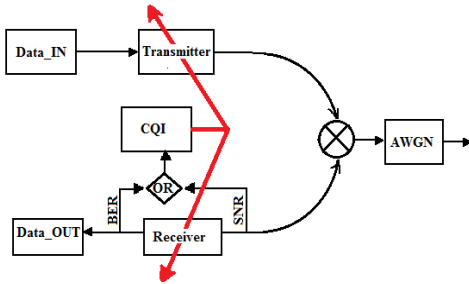


Figure 8. Relation Between SDR and CE in a CR Transceiver

The encoding process was not taken into account when designing this function, when used; it will significantly enhance the throughput. From the SDR function the effect of encoding is not on the scope. From the cognitive engine perspective, the improvement made by encoding is directly reflected on the BER; therefore, it's not a cognitive engine matter.

B. Real world Implementation of the AMC

CE successfully tracks the SNR and selects the suitable modulation order in order to maintain the desired BER specified by the application used. As in Figure 7 A, the scattered SNR levels are generated to test the CE in the worst case scenario. However, for real world implementations, a distance based slow-varying SNR model is adopted to analyze the performance of the CE comparing to the AMC developed using hardware approaches.

Fortunately, this slow-varying assumption allows the receiver part of the communication system to anticipate the channel conditions for the next transmission interval in terms of CQI. Since this knowledge can only be gained from projections using past CQI, so AMC based systems can operate more efficiently under this assumption.

When multipath Fading or Doppler effects take place because of the multipath signals or the relatively high speed mobility of the CPE, the communication system needs more time to estimate the SNR level for the next transmission interval. However, when the transmitted packets are coherent (e.g. VoIP) the system has to tolerate some errors to reduce delay and jitter. This inaccuracy is crucial for AMC system implementation, since a poor system performance will result if the channel estimate is obsolete and/or inaccurate at the time of transmission.

For instance, a handoff scenario is assumed here to verify the system performance in real world, which demonstrates a realized scenario with a channel variation from good, bad, then again good SNR level. This scenario just affects the channel condition faced by the CE, and has nothing to do with the main CE mechanism. Obviously, the functionality of the CE will remain the same as in Figure 7; however, in this scenario the simulation environment tends to behave as in real world operation.

C. Performance Analysis of the AMC Cognitive Engine

Table 1 represents the percentage of packets transmitted using the modulation order selected by the CE function when the user specified the desired BER and according to the instantaneous channel conditions. The "Low SNR" column gives the percentage of the packets that didn't transmit because of the low SNR level; this means that there is no modulation order sufficient to maintain the desired BER for the given channel condition. In this scenario the channel condition represented in SNR follows a handover procedure with SNR varying gradually according to the effect of the long-term fading while the mobile CPE moves to another cell area. CE maintains the desired BER in the varying channel condition by altering the modulation order.

TABLE I. PERCENTAGE OF MODULATION ORDER USED

Modulation Order (%)	BPSK	QPSK	16QAM	64QAM	Low SNR
10^{-1}	16	26	22	21	15
10^{-2}	14	23	22	10	31
10^{-3}	10	25	22	3	40

The average system efficiency can be calculated for the designed function using the method specified in [8] and is defined as;

$$\bar{\eta}_{ave} = \frac{N_{64} \times 8 \times 4.5 + N_{16} \times 12 \times 3 + N_4 \times 24 \times 1.5 + N_2 \times 72 \times 0.5}{N_{64} \times 8 + N_{16} \times 12 + N_4 \times 24 + N_2 \times 72}$$

where N represents the percentage of the MCS used, and the efficiency coefficients for each modulation order are taken from Table II.

TABLE II. EFFICIENCIES OF MCSs

Modulation	Efficiency
BPSK	0.5
QPSK	1.5
16QAM	3
64QAM	4.5

The average efficiency of the transmission using the CE function is shown in Table III when the user specified the desired BER and assuming slow varying channel condition.

Table III shows the direct relationship between the desired BER and average efficiency. Another important observation is that the average efficiency is always better using low modulation orders (BPSK or QPSK) in most times. Obviously, using higher modulation orders (i.e.

16QAM or 64QAM) is better in terms of spectral efficiency. However, due to their susceptibility to errors, it is not recommended to utilize higher modulation orders for the entire transmission when the CPE faces low SNR.

TABLE III. AVERAGE EFFICIENCY OF THE AMC USING CE

BER	Average Efficiency
10^{-1}	2.075
10^{-2}	1.525
10^{-3}	1.22

V. CONCLUSIONS

Following the ever changing communication protocols is one of the major expenditures for the networks operators. However, reconfigurable platforms will contribute to minimize this cost. Moreover, CR can offer more than just providing the same hardware functionalities in software. To discover what CR can offer beyond that, simulations test beds may trigger some thoughts [9]. One of the main pros of the CR, that is enabling system engineers to focus on applications of interest rather than low-level details, which reduce design effort, provides higher reliability, and allows easy deployment on different target platforms.

In this paper an implementation of the CE performed as a module to work with a wide range of MCS profiles in AMC based communication system for BWA based on CR approaches is presented. A unified test bench for AMC function using CR approaches is performed using computer simulation, the compatibility between the SDR and the CE part and its benefit to the average efficiency is verified. Both functionality and system performance are verified with the legacy hardware approach.

REFERENCES

- [1] Sami Salih, and Mamoun Suliman, "Implementation of adaptive modulation and coding techniques using Matlab", IEEE, ELMAR, 2011.
- [2] IEEE 802.16-2006: "IEEE Standard for Local and Metropolitan Area Networks - Part 16: Air Interface for Fixed Broadband Wireless Access Systems".
- [3] Dania Marabissi and others, "Efficient Adaptive Modulation and Coding techniques for WiMAX systems", ICC 2008 proceedings.
- [4] Otuski, Nobuaki; Yoshizawa, Shingo; Miyanaga, Yoshikazu, "A New Adaptive Modulation and Coding Applied on a 400-Mbps OFDM Wireless Communication System", 2006. ISIT '06.
- [5] Hazem Shatila, Mohamed Khedr, Jeffrey H. Reed, "Adaptive Modulation and Coding for WiMAX Systems with Vague Channel State Information using Cognitive Radio", SPECTS, 2010.
- [6] Kaushik R. Chowdhury and Tommaso Melodia, "Platforms and Testbeds for Experimental Evaluation of Cognitive Ad Hoc Networks", IEEE Communication Magazine, Vol. 48, No. 9, September 2010.
- [7] Paul D. Sutton, "Iris: An Architecture for Cognitive Radio Networking Testbeds", IEEE Communication Magazine, Vol. 48, No. 9, September, 2010.
- [8] Mohammad Reza, Hassan Taheri, "Mobile WiMAX Capacity Estimation in Various Conditions", 18th Iranian Conference on Electrical Engineering (ICEE), 2010.
- [9] J. Lotze, S.A. Fahmy, J. Noguera and L.E. Doyle, "A Model-Based Approach to Cognitive Radio Design", IEEE Selected Areas in Communications, Vol. 29, No. 2, pp. 455-468, February 2011.
- [10] Zhenyu Zhang, Xiaoyao Xie, "Intelligent cognitive radio: Research on learning and evaluation of CR based on Neural Network", ICIT 2007.
- [11] Colson, N., and others, "Autonomous Decision Making Process for the Dynamic Reconfiguration of Cognitive Radios", Computer Communications and Networks (ICCN), 2008.

SaPM: Switch-aware Process Mapping Model for Parallel Computing

Yufei Lin, Yuhua Tang, Xinhai Xu
State Key Laboratory of High Performance Computing,
School of Computer Science, National University of Defense Technology,
ChangSha, Hunan, 410073, China.
{linyufei, tangyuhua, xuxinhai}@nudt.edu.cn

Abstract—The problem of assigning processes of a parallel program to processors of a parallel system, namely process mapping problem, has major impact on the resulting performance. The previous works only concern about the computation and communication cost, while ignoring the process switching cost. For a large-scale and communication-intensive application, more than one process may be placed onto a same processor. When the application is running, it will take long time to switch the processes. In this paper, we propose and introduce SwitchCAL, a method to calculate the process switching cost inside a processor, into the process mapping problem. Then we propose a switch-aware process mapping model – SaPM, and integrate the SaPM model with existing process mapping algorithm to obtain optimized process mapping scheme. The experiments show that our approach has high effectiveness.

Index Terms—process mapping; process switching; graph mapping; MPI

I. INTRODUCTION

The problem of assigning processes of a parallel program to processors of a parallel system, namely process mapping problem, has major impact on the resulting performance. For the system designers, the performance they concern about is the execution time of a parallel program. Therefore the minimax criterion [1] of the process mapping problem is widely used. The goal of this criterion is to minimize the maximum workload per processor, which is defined as the total cost due to the computation and communication of all the processes mapped to it (workload per processor = computation cost + communication cost) [2].

However, when the application has a larger scale than the parallel machine, such as the execution-driven simulator Bigsim [3], more than one processes will be mapped onto a same processor. Furthermore, if the application is communication-intensive, which means that the process on the same processor are switched frequently, the switching cost will badly influence the performance and it should not be ignored. But all the existing works have ignored the influence of the process switching due to the communications among the processes in the same processor.

In this paper, we integrate for the first time the influence of process switching into the process mapping problem. Our main contributions are:

- Propose a method to calculate the process switching cost inside a processor – SwitchCAL.

- Introduce SwitchCAL into the process mapping problem, and propose a switch-aware process mapping model – SaPM.
- Integrate the SaPM model with existing process mapping algorithm to obtain optimized process mapping.
- Demonstrate the effectiveness of our method by experiments.

The rest of the paper is organized as follows. Section II proposes the SaPM model. Section III introduce the SwitchCAL method and the optimized objective of SaPM model. The experiments are demonstrated in Section IV. Section V reviews the related work. Finally, Section VI summarizes the paper.

II. SAPM MODEL

The problem of process mapping can be formalized to a graph mapping problem which finds the optimized mapping between the Task Interaction Graph (TIG) of applications and the Network Topology Graph (NTG) of the underlying parallel computer systems. Therefore, the SaPM model is composed of three parts: defining and conducting TIG and NTG, deciding the optimized objective of SaPM model, designing the mapping algorithm. Shown as in Fig. 1, with the TIG and NTG extracted, the graph mapping algorithm will output the optimized mapping result by designing a proper optimized objective.

A. TIG and NTG

For a parallel program, a formal definition of TIG is $G_P(N_P, E_P, W_P, C_P)$, where

- N_P is the set of the processes.
- E_P is the set of edges $\{(p_i, p_j) | p_i, p_j \in N_P\}$ such that there is communication between p_i and p_j .
- $w_{p_i} \in W_P$ associated with each node is the computation time of process $p_i \in N_P$.
- $cp_{p_i, p_j} \in W_P$ is the weight of edge $p_i, p_j, cp_{p_i, p_j} = \{(Count_{p_i, p_j}, Volume_{p_i, p_j}) | p_i, p_j \in N_P\}$, where $Count_{p_i, p_j}$ and $Volume_{p_i, p_j}$ are respectively the total message count and size between communicating processes p_i and p_j .

TIG describes the communication characteristics of an parallel application, and is obtained by using the profiling tools, such as Intel Trace Collector [4] or VampirTrace [5].

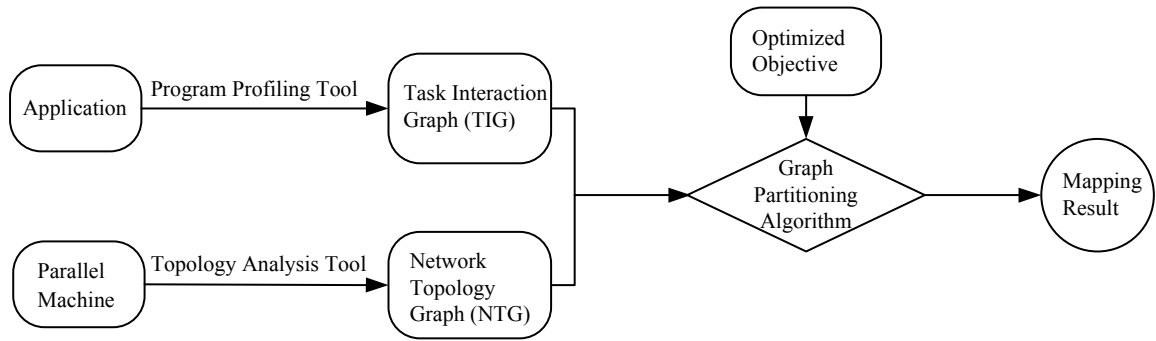


Fig. 1: The workflow of SaPM model.

For a parallel machine, the definition of NTG G_M is $G_M(N_M, C_M)$, where

- N_M is the set of processing cores.
- $cm_{m_s, m_t} \in C_M$ is the set of network characteristics between $m_s \in N_M$ and $m_t \in N_M$. $cm_{m_s, m_t} = \{(L_{m_s, m_t}, B_{m_s, m_t}) | m_s, m_t \in N_M\}$ where L_{m_s, m_t} and B_{m_s, m_t} respectively represent the latency and bandwidth between m_s and m_t .

NTG reflects the network parameters of the target machine, and is conducted by running ping-pong program [6], [7], logmp benchmark tool [8] or supplied by users [9].

B. Optimized Objective

The optimized objective of SaPM model is to minimize the maximum total cost due to the computation, communication and process switching per processor:

$$\min_{m_s \in N_M} (Cost_{comp}(m_s) + Cost_{comm}(m_s) + Cost_{swi}(m_s))$$

where $Cost_{comp}(m_s)$, $Cost_{comm}(m_s)$ and $Cost_{swi}(m_s)$ are respectively the computation cost, communication cost and process switching cost of processor m_s .

Define function $f : N_P \rightarrow N_M$ as the mapping function: if process $p_i \in N_P$ is mapped onto processor $m_s \in N_M$, then denote $f(p_i) = m_s$.

The computation cost of processor m_s is

$$Cost_{comp}(m_s) = \sum_{p_i | f(p_i) \in m_s} w(p_i)$$

The communication cost of processor m_s is the cost of the communications between the processes mapped on it and the processes mapped on the other processors:

$$Cost_{comm}(m_s) = \sum_{p_i | f(p_i) = m_s} \sum_{p_j | f(p_j) \neq m_s} (Count_{p_i, p_j} L_{m_s, m_t} + \frac{Volume_{p_i, p_j}}{B_{m_s, m_t}})$$

The methods of calculating $Cost_{swi}(m_s)$ is difficult, and we will discuss it in Section III.

C. Graph Mapping Algorithm

Lots of graph mapping algorithms are studied to figure out the process mapping problem under minimax criterion, such as [10] and [11], etc. Both of them can be used as the basis to solve the graph mapping problem with our objective function. Input TIG and NTG into a graph mapping algorithm, then it can output the optimized mapping result matching the optimized objective.

III. SWITCHCAL

In this section, we will introduce SwitchCAL, a method to calculate the process switching cost inside a processor.

Nowadays, most large scale parallel codes are written using MPI, which has become the de facto standard for parallel computing[12]. MPI will switch the communication protocol according to the size of the message to be transferred. There are two kinds of communication protocols: asynchronous protocol is used for small messages while synchronous protocol for large messages [13], shown as in Fig. 2.

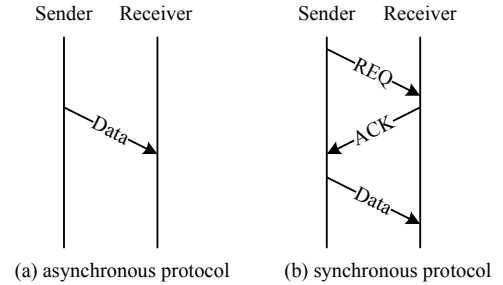


Fig. 2: The communication protocol.

For small messages, asynchronous protocol sends the data message immediately to the buffer of the receiver directly. For long messages, synchronous protocol needs a handshake between the sender and the receiver via REQ and ACK messages before the data message is sent to the receiver. So for small messages, a process will not be switched off. Therefore, we first calculate the median message size between the communicating processes and compare it with a *Threshold* to determine whether it is dominated by large messages or small ones. If the median message size is below the Threshold,

the process switching cost is ignored. Then if process p_i and p_j are mapped onto a same processor m_s , then the switching cost due to process p_i and p_j is:

$$cost_{swi}(p_i, p_j) = \begin{cases} Count_{p_i, p_j} \times O_s, & \text{if } \frac{Volume_{p_i, p_j}}{Count_{p_i, p_j}} \geq Threshold \\ 0, & \text{otherwise} \end{cases}$$

where O_s is the time cost per switching.

Therefore, we can derive the total switching cost on processor m_s :

$$Cost_{swi}(m_s) = \sum_{p_i, p_j | f(p_i)=m_s, f(p_j)=m_s} cost_{swi}(p_i, p_j)$$

IV. EXPERIMENTS

It is not difficult to find out that when the message volume are small, our SaPM model will degenerate to the existing models, so we choose MG (Class A), LU (Class A), CG (Class A) and BT (Class A), whose message volumes are large, in NPB3.3-MPI benchmarks to evaluate SaPM. MG computes an approximate solution to the discrete Poisson problem using four iterations of the V-cycle multi-grid algorithm on a $n \times n \times n$ grid with periodic boundary conditions [12]. The LU benchmark solves the 3D Navier-Stokes equation by using a Successive Over-Relaxation (SSOR) algorithm [12]. CG tests the performance of the system for unstructured grid computations which by their nature require irregular long distance communications [14]. BT solves three sets of uncoupled systems of equations, first in the x , then in the y , and finally in the z direction [14]. For each benchmark, we start four processes. The TIGs of MG, LU, CG and BT are shown in Fig. 3.

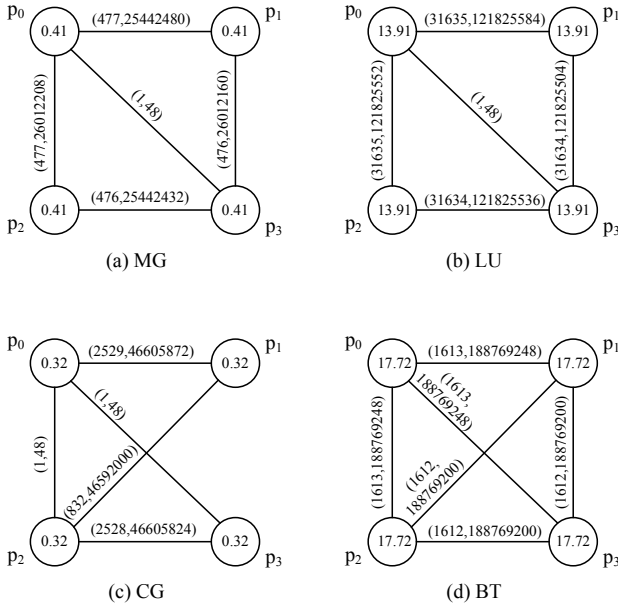


Fig. 3: The TIGs of MG, LU, CG and BT.

Our experimental platform is two processing node, each of which is equipped with two 2.93G 6-core Intel Xeon X5670

CPUs and 24GB RAM. Because we will map more than one process onto a processor, we use just one core of each node. The network latency is $4.06 \times 10^{-5} s$, and the network bandwidth is $1.14 \times 10^7 bytes/s$.

Then we compare the mapping scheme obtained by SaPM which considers the process switching cost to the scheme obtained without considering process switching. For MG, without considering process switching, process 0 and 1 are mapped together, while considering process switching by using SaPM, process 0 and 2 are mapped together. For LU, without considering process switching, process 0 and 1 are mapped together, while considering process switching by using SaPM, process 0 and 3 are mapped together. For CG, without considering process switching, process 0 and 1 are mapped together, while considering process switching by using SaPM, process 0 and 3 are mapped together. For BT, without considering process switching, process 0 and 1 are mapped together, while considering process switching by using SaPM, process 0 and 2 are mapped together.

The execution time of MG, LU, CG and BT under different schemes is shown in Fig. 4.

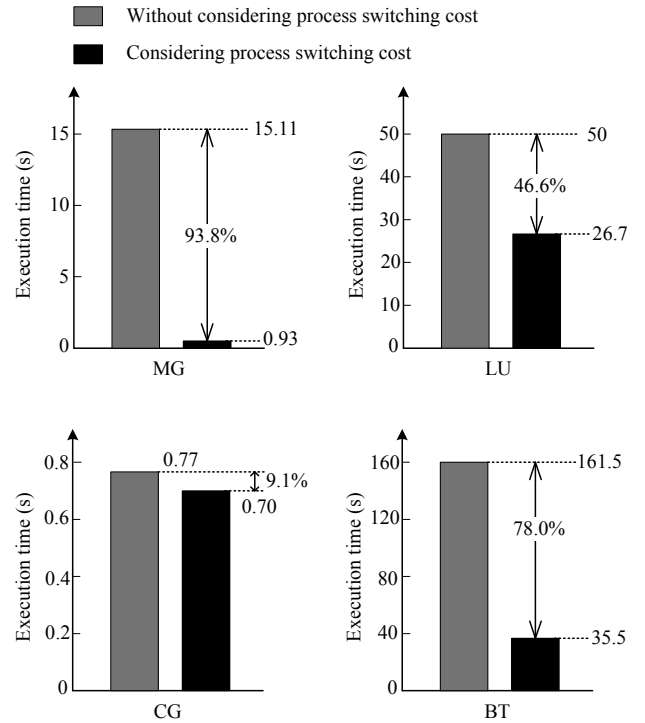


Fig. 4: The execution time of MG, LU, CG and BT under different mapping schemes

From Fig. 4, we can see that the execution time is reduced by 93.8%, 46.6%, 9.1% and 78.0%. The results not only show the effectiveness of SaPM, but also prove that the process switching cost has high influence on the execution time.

V. RELATED WORK

Besides the Task Interaction Graph (TIG), there are other descriptions for parallel applications, such as Task Precedence

Graph (TPG) [15] and Temporal Task Interaction Graph (TTIG) [16]. TPG is a directed acyclic graph. The directed edge indicates the communication source and destination. TTIG is a new task graph model, which considers concurrency of the tasks.

As the mapping problem is NP-complete [17], many heuristics-based solutions have been proposed, including graph theoretic, mathematical programming, state-space search, probabilistic and randomized optimization methods [18]. Concretely, there are simulated annealing [19], graph contraction [20], genetic algorithms [21], and mincut bipartitioning [22], etc. Our algorithm is based on the k-way graph partitioning algorithm proposed in [6], which is proved to be a more effective graph partitioning algorithm for multi-clusters than previous works.

There are also optimization methods focusing on some specific parallel machines or topologies, such as [23] for Cray T3E, [24] for Bluegene/L, [25] for mesh and [26] for hierarchical NUMA architecture, etc.

Our approach, to the best of our knowledge, is the first work to obtain optimized process mapping by taking the process switching cost into consideration.

VI. CONCLUSION

Process mapping is a fundamental problem in parallel computing. In this paper, we take the process switching cost into consideration. We propose a method to calculate the process switching cost inside a processor, SwitchCAL, at first. Next, we introduce SwitchCAL into our switch-aware process mapping model – SaPM. Then integrate SaPM with existing process mapping algorithm proposed in [6] to obtain optimized process mapping scheme. At last, we prove the high effectiveness of our method by experiments. For MG, LU, CG and BT in NPB benchmarks, the execution time is reduced by 93.8%, 46.6%, 9.1% and 78.0% respectively.

ACKNOWLEDGMENT

This work is supported by the National Natural Science Foundation of China (NSFC) No.60921062.

REFERENCES

- [1] C.-C. Shen and W.-H. Tsai, "A graph matching approach to optimal task assignment in distributed computing systems using a minimax criterion," *IEEE Trans. Comput.*, vol. 34, no. 3, pp. 197–203, Mar. 1985.
- [2] C. Roig, A. Ripoll, M. A. Senar, F. Guirado, and E. Luque, "Exploiting knowledge of temporal behaviour in parallel programs for improving distributed mapping," in *Proceedings from the 6th International Euro-Par Conference on Parallel Processing*, ser. Euro-Par '00. London, UK, UK: Springer-Verlag, 2000, pp. 262–271.
- [3] G. Zheng, G. Kakulapati, and L. Kale, "Bigsim: a parallel simulator for performance prediction of extremely large parallel machines," in *Parallel and Distributed Processing Symposium, 2004. Proceedings. 18th International*, april 2004, p. 78.
- [4] Website, http://software.intel.com/sites/products/documentation/hpc/itac/itc_reference_guide.pdf.
- [5] Website, <http://www.vampir.eu/>.
- [6] H. Chen, W. Chen, J. Huang, B. Robert, and H. Kuhn, "Mpipp: an automatic profile-guided parallel process placement toolset for smp clusters and multiclusters," in *Proceedings of the 20th annual international conference on Supercomputing*, ser. ICS '06. New York, NY, USA: ACM, 2006, pp. 353–360.
- [7] J. Zhang, J. Zhai, W. Chen, and W. Zheng, "Process mapping for mpi collective communications," in *Proceedings of the 15th International Euro-Par Conference on Parallel Processing*, ser. Euro-Par '09. Berlin, Heidelberg: Springer-Verlag, 2009, pp. 81–92.
- [8] T. Kielmann, H. E. Bal, and K. Verstoep, "Fast measurement of logp parameters for message passing platforms," in *Proceedings of the 15 IPDPS 2000 Workshops on Parallel and Distributed Processing*, ser. IPDPS '00. London, UK, UK: Springer-Verlag, 2000, pp. 1176–1183.
- [9] A. Pant and H. Jafri, "Communicating efficiently on cluster based grids with mpich-vmi," in *Cluster Computing, 2004 IEEE International Conference on*, sept. 2004, pp. 23 – 33.
- [10] M. Kafil and I. Ahmad, "Optimal task assignment in heterogeneous distributed computing systems," *Concurrency, IEEE*, vol. 6, no. 3, pp. 42–50, jul-sep 1998.
- [11] M. A. Senar, A. Ripoll, A. Cortés, and E. Luque, "Clustering and reassignment-based mapping strategy for message-passing architectures," *J. Syst. Archit.*, vol. 48, no. 8-10, pp. 267–283, Mar. 2003.
- [12] C. Coarfa, "Portable high performance and scalability of partitioned global address space languages," Ph.D. dissertation, RICE UNIVERSITY, 2007.
- [13] W. Gropp, E. Lusk, N. Doss, and A. Skjellum, "A high-performance, portable implementation of the mpi message passing interface standard," *Parallel Comput.*, vol. 22, pp. 789–828, September 1996.
- [14] A. Afsahi, "Design and evaluation of communication latency hiding/reduction techniques for message-passing environments," Ph.D. dissertation, University of Victoria, 2000.
- [15] Y.-K. Kwok and I. Ahmad, "Dynamic critical-path scheduling: an effective technique for allocating task graphs to multiprocessors," *Parallel and Distributed Systems, IEEE Transactions on*, vol. 7, no. 5, pp. 506–521, may 1996.
- [16] C. Roig, A. Ripoll, and F. Guirado, "A new task graph model for mapping message passing applications," *IEEE Trans. Parallel Distrib. Syst.*, vol. 18, no. 12, pp. 1740–1753, Dec. 2007.
- [17] S. H. Bokhari, "On the mapping problem," *IEEE Transactions on Computers*, vol. 30, pp. 207–214, 1981.
- [18] B. Ucar, C. Aykanat, K. Kaya, and M. Ikinici, "Task assignment in heterogeneous computing systems," *J. Parallel Distrib. Comput.*, vol. 66, no. 1, pp. 32–46, Jan. 2006.
- [19] S. Bollinger and S. Midkiff, "Heuristic technique for processor and link assignment in multicomputers," *Computers, IEEE Transactions on*, vol. 40, no. 3, pp. 325–333, mar 1991.
- [20] F. Berman and L. Snyder, "On mapping parallel algorithms into parallel architectures," *J. Parallel Distrib. Comput.*, vol. 4, no. 5, pp. 439–458, Oct. 1987.
- [21] P. Neuhaus, "Solving the mapping problem - experiences with a genetic algorithm," in *Proceedings of the 1st Workshop on Parallel Problem Solving from Nature*, ser. PPSN I. London, UK, UK: Springer-Verlag, 1991, pp. 170–175.
- [22] F. Ercal, J. Ramanujam, and P. Sadayappan, "Task allocation onto a hypercube by recursive mincut bipartitioning," in *Proceedings of the third conference on Hypercube concurrent computers and applications: Architecture, software, computer systems, and general issues - Volume 1*, ser. C3P. New York, NY, USA: ACM, 1988, pp. 210–221.
- [23] E. Huedo, M. Prieto, I. M. Llorente, and F. Tirado, "Impact of pe mapping on cray t3e message-passing performance," in *Proceedings from the 6th International Euro-Par Conference on Parallel Processing*, ser. Euro-Par '00. London, UK, UK: Springer-Verlag, 2000, pp. 199–207.
- [24] G. Bhanot, A. Gara, P. Heidelberger, E. Lawless, J. C. Sexton, and R. Walkup, "Optimizing task layout on the blue gene/l supercomputer," *IBM Journal of Research and Development*, vol. 49, no. 2.3, pp. 489–500, march 2005.
- [25] A. Bhatele and L. V. Kale, "Benefits of topology aware mapping for mesh interconnects," *Parallel Processing Letters*, vol. 18, pp. 549–566, 2008.
- [26] E. Jeannot and G. Mercier, "Near-optimal placement of mpi processes on hierarchical numa architectures," in *Proceedings of the 16th international Euro-Par conference on Parallel processing: Part II*, ser. Euro-Par '10. Berlin, Heidelberg: Springer-Verlag, 2010, pp. 199–210.

New Procedure for Discrimination of Model Parameter and Noise Variance Changes

Theodor D. Popescu

National Institute for Research and Development in Informatics
8-10 Averescu Avenue, 011455 Bucharest
Romania

Abstract—The change detection and diagnosis methods have gained considerable attention in scientific research and appears to be the central issue in various application areas. These applications need some robust change detection schemes to work well and separate the changes in the experimental conditions from the real changes in the system, especially for systems with arbitrary and non-stationary known or unknown inputs. The objective of the paper is to develop a such kind of change detection and diagnosis scheme, able to discriminate the model parameter and noise variance changes. Finally, we include some Monte-Carlo simulations for change detection in a second order FIR model and experimental results obtained in analysis of seismic signals, using the proposed approach.

I. INTRODUCTION

The problem of change detection and diagnosis has gained considerable attention during the last two decades in a research context and appears to be the central issue in various application areas such as: speech processing, image processing, analysis of biomedical signals, signal processing in cars, digital data transmission systems, underwater acoustics, geophysics, failure detection in controlled systems (aeronautics, chemical and nuclear processes, event detection of incident on freeways, leak detection for pipelines), econometrics, etc.

From statistical point of view, change detection tries to identify changes in the probability distribution of a stochastic process. In general, the problem involves both detecting whether or not a change has occurred, or whether several changes might have occurred, and identifying the times of any such changes. Specific applications may be concerned with changes in the mean, variance, correlation, or spectral density of the process.

Deeper interactions between the control, signal processing, and statistical communities recently contributed to the insight in the change detection problem in a significant way. The theory has been used in many successful applications. Among these can be mentioned applications in mechanical engineering, industrial process monitoring, civil infrastructure, medical diagnosis and treatment, speech segmentation, underwater sensing, video surveillance, and driver assistance systems, among others.

Almost all change detection schemes make use of the assumption that the system itself can be accurately described by a linear finite dimensional model. This leads to the conclusion that arbitrary small changes can be detected for sufficient data

records. In practice, there are several problems associated with the usual test procedures:

- 1) When the system is more complex than the finite dimensional model structure, the parameter estimate will still converge, but to a value that now depends upon the experimental conditions. The calculations of the covariance matrix will be much more involved. If the experimental conditions are different, the difference between the parameter vector resulted for a normal operating data set and the parameter vector for an fault operating data set has no longer zero mean value and the common tests will fail, since the algorithms can not separate changes due to variations in experimental conditions and the real changes in the investigated system.
- 2) If the parameter vector is a black-box parameter vector (the coefficients of the model polynomial are the most common choice), it may be difficult to get insight into how variations in the behavior of the system reflect in variations in parameter vector. Very different choices of parameter vector can give good approximations of the same system, which cause obvious problems in change detection schemes.

All the problems mentioned above point out the need for some robust change detection schemes to work well and separate the changes in the experimental conditions from the real changes in the system, especially for systems with arbitrary and non-stationary known or unknown inputs. Examples of such systems are vibration structures: buildings subject to earthquakes, off-shore platforms subject to wave action, bridges subject to wind or earthquakes, mechanical systems subject to fluid interaction. The objective of the present paper is to develop a such change detection scheme that may be able to separate the changes in model parameters and noise variance, starting from the classical change detection schemes based on sliding windows approach and likelihood techniques.

An original approach for change detection and diagnosis in modal characteristics of non-stationary signals for unmeasured natural excitation and application of this technique in vibration monitoring is presented in [1]. This is based upon the use of instrumental statistics and a statistical approach for detection. A practical application of this approach is given in [2]. Another general approach for measured inputs, suggested in [3], makes use of some tests developed in frequency domain. This

approach is especially recommended for diagnosis purposes. The major part of all change detection schemes seems to be more robust and insightful to work in frequency domain. An application of this approach for a structural system is presented in [4].

The problem formulation, discussed in the paper, assumes the off-line or batch-wise data processing, although the solution is sequential in data and an on-line data processing can be used. The change detection model is the simplest possible extension of linear regression models to data with abruptly changing properties. It is assumed that the data can be described by one linear regression model within each segment with distinct parameter vector and noise variance.

The outline of this paper is as follows. In Section 2, the conceptual description of a new algorithm able to discriminate the model parameter and noise variance changes is given. Finally, in Section 3, we include some Monte-Carlo simulations for change detection in a second order FIR (Finite Impulse Response) model and experimental results obtained in analysis of seismic signals, using the proposed approach.

II. DETECTION AND DISCRIMINATION OF PARAMETERS AND VARIANCE CHANGES

This algorithm makes use of the concepts classical algorithms in this field, but its novelty consists in the fact that it is able to discriminate between the changes appeared in parameters and noise variance at the change instant, t_0 . This is of great practical importance in many applications, where the classical algorithms are unable to discriminate between the changes in the system and in the environment. The detection and isolation algorithm in this case is based on likelihood. The following signal model can be used, [5]:

$$y_t = \begin{cases} \phi_t^T \theta_0 + e_t, & E[e_t] = \lambda_0 R_t & \text{if } t \leq t_0 \\ \phi_t^T \theta_1 + e_t, & E[e_t] = \lambda_1 R_t & \text{if } t > t_0 \end{cases} \quad (1)$$

where R_t is a known time-varying noise (co-)variance, and λ is either a scaling of the noise variance or the variance itself ($R_t = 1$). Neither θ_0 , θ_1 , λ_0 or λ_1 are known. The following hypotheses are used:

$$\begin{aligned} H_0: & \theta_0 = \theta_1 \quad \text{and} \quad \lambda_0 = \lambda_1 \\ H_1: & \theta_0 \neq \theta_1 \quad \text{and} \quad \lambda_0 = \lambda_1 \\ H_2: & \theta_0 = \theta_1 \quad \text{and} \quad \lambda_0 \neq \lambda_1 \end{aligned} \quad (2)$$

The sufficient statistics from the filters are given below:

Data	$y_1, y_2, \dots, y_{t-L},$	y_{t-L+1}, \dots, y_t	
Model	M_0	M_1	
Time interval	T_0	T_1	
RLS quantities	$\hat{\theta}_0, P_0$	$\hat{\theta}_1, P_1$	(3)
Loss function	V_0	V_1	
Number of data	$n_0 = t - L$	$n_1 = L$	

where $P_j, j = 0, 1$ denotes the covariance of the parameter estimate achieved from the RLS algorithm. The loss functions are defined by

$$V_j(\theta) = \sum_{k \in T_j} (y_k - \phi_k^T \theta)^T (\lambda_j R_k)^{-1} (y_k - \phi_k^T \theta), \quad j = 0, 1. \quad (4)$$

It makes sense to compute $V_1(\hat{\theta}_0)$ to test how the first model performs on the new data set. The maximum likelihood approach is stated in the slightly more general maximum *a posteriori* approach, where the prior probabilities q_i for each hypothesis can be incorporated. The exact *a posteriori* probabilities

$$l_i = 2 \log p(H_i | y_1, y_2, \dots, y_t), \quad i = 0, 1, 2 \quad (5)$$

are derived by Gustafsson [6]. Assuming that $H_i, i = 0, 1, 2$ is Bernoulli distributed with probability q_i , i.e.

$$H_i = \begin{cases} \text{does not hold,} & \text{with probability } 1 - q_i \\ \text{hold,} & \text{with probability } q_i \end{cases} \quad (6)$$

$\log p(H_i)$ is given by

$$\begin{aligned} \log p(H_i) &= \log(q_i^2 (1 - q_i)^{n_0 + n_1 - 2}) \\ &= 2 \log(q_i) + (n_0 + n_1 - 2) \log(1 - q_i), \quad (7) \\ & \quad i = 0, 1, 2. \end{aligned}$$

For the signal model (1) where $e \in N(0, \lambda)$, the prior distribution for λ can be taken as inverse Wishart distribution (or gamma distribution in scalar case). The inverse Wishart distribution has two parameters, m and σ , and is denoted by $W^{-1}(m, \sigma)$. Its probability density function is given by

$$p(\lambda) = \frac{\sigma^{m/2} e^{-\frac{\sigma}{2\lambda}}}{2^{m/2} \Gamma(m/2) \lambda^{(m+2)/2}} \quad (8)$$

The expected mean value of λ is

$$E(\lambda) = \frac{\sigma}{m - 2} \quad (9)$$

and the variance is given by

$$\text{Var}(\lambda) = \frac{2\sigma^2}{(m - 2)^2 (m - 4)} \quad (10)$$

The mean value and the noise variance are design parameters and from these the Wishart parameter m and σ can be computed. For the signal model (1) and the hypothesis given in (2), let the prior for λ as in (8) and the prior for the parameter vector be $\theta \in N(0, P_0)$. With the loss function (10) and the least squares estimation, the *a posteriori* probabilities are approximately given by (see [6]):

$$\begin{aligned} l_0 &\approx (n_0 + n_1 - 2 + m) \log \left(\frac{V_0(\hat{\theta}_0) + V_1(\hat{\theta}_1) + \sigma}{n_0 + n_1 - 4} \right) \\ &\quad + \log \det(P_0^{-1} + P_1^{-1}) + 2 \log(q_0), \end{aligned} \quad (11)$$

TABLE I
MODEL PARAMETERS AND λ VALUES

Time	1-50	51-100	101-150	151-200	201-250
θ_1	1.6	-1.6	-1.6	-1.6	1.6
θ_2	0.64	0.64	0.64	0.64	0.64
λ	0.5	0.5	5.	0.5	0.5

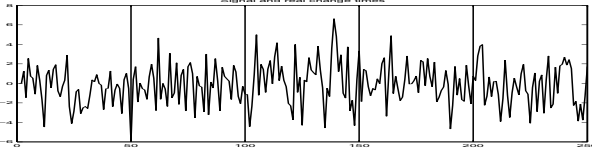


Fig. 1. The signal and the real change instants

$$l_1 \approx (n_0 + n_1 - 2 + m) \log \left(\frac{V_0(\hat{\theta}_0) + V_1(\hat{\theta}_1) + \sigma}{n_0 + n_1 - 4} \right) - \log \det P_0 - \log \det P_1 + 2 \log(q_1), \quad (12)$$

$$l_2 \approx (n_0 - 2 + m) \log \left(\frac{V_0(\hat{\theta}_0) + \sigma}{n_0 - 4} \right) + (n_1 - 2 + m) \log \left(\frac{V_1(\hat{\theta}_1) + \sigma}{n_1 - 4} \right) - 2 \log \det P_0 + 2 \log(q_2). \quad (13)$$

The last three equations are used in decision making concerning one of the three hypotheses presented above.

III. EXPERIMENTAL RESULTS

A. Simulation results

The presented results are obtained by Monte Carlo simulation for a second order FIR model with 4 change times. The following model structure was used:

$$y_t = \theta_1 * u_{t-1} + \theta_2 * u_{t-2} + e_t \quad (14)$$

where u_t and e_t are random sequences of zero mean and variance $E[u_t] = 1$ and $E[e_t] = \lambda$, respectively. The model parameters and λ values for the first experiment are given in the Table 1. The changes produced, for the parameter θ_1 (the parameter θ_2 is kept constant during the experiment), at the instants 51 and 201, and for the scaling factor of the noise variance, λ , at the instants 101 and 151.

The signal resulted by simulation, for a noise realization of the random sequence, e_t , and the real change instants are given in Fig. 1.

The Monte Carlo simulation consisted in simulation of the model given in (14), with the model parameters and λ values given in Table 1, for 1000 realizations of the random sequence, e_t , and in applying of the algorithm for diagnosis of parameters and noise variance changes. For each of the 1000 experiments resulted the parameter estimates and noise variance estimate.

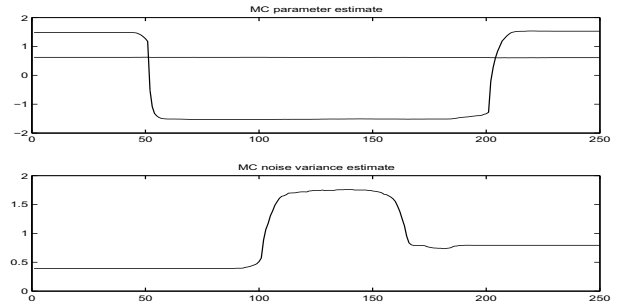


Fig. 2. Model parameter and noise variance estimates obtained in Monte Carlo simulation

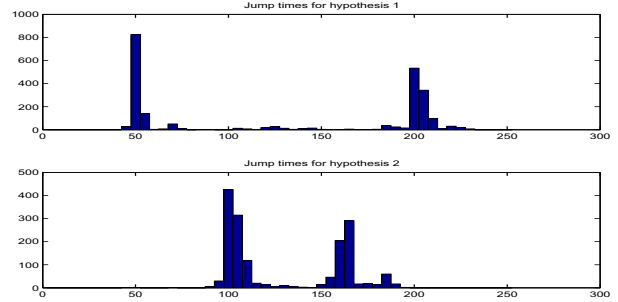


Fig. 3. The histogram for change detection instants in model parameters and noise variance

These values have been averaged and are presented in Fig. 2. It can be noted very good values for model parameter estimates, according with the real values, given in Table 1. Concerning the noise variance estimate, the changes in this value are detected at the real instants, but the estimates are not so good as parameter estimates. This can explained by the reduced values of the noise.

From the histogram for change detection instants, given in Fig. 3, it can be noted the great rate of real change detection instants, especially for model parameters. Also, it points out some delays in detecting of the real change instants, and a reduced false alarms. The results depend to a great extend of the signal/noise ratio.

A similar experiment was performed for the model parameters and noise variance, λ , values given in Table 2.

The signal resulted by simulation for the same FIR process, a noise realization of the random sequence, e_t , and the real change instants are given in Fig. 4, while in Fig. 5 the model parameter and noise variance estimates obtained in a Monte Carlo simulation for 1000 noise realization are presented. The resulted histogram for change detection instants is given in

TABLE II
MODEL PARAMETERS AND λ VALUES

Time	1-50	51-100	101-150	151-200	201-250
θ_1	1.6	-1.6	-1.6	-1.6	1.6
θ_2	0.64	0.64	0.64	0.64	0.64
λ	1.	1.	10.	1.	1.

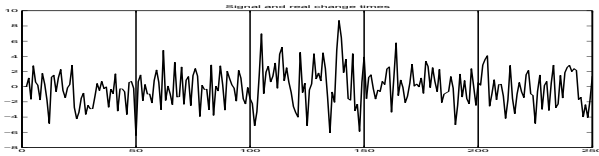


Fig. 4. The signal and the real change instants

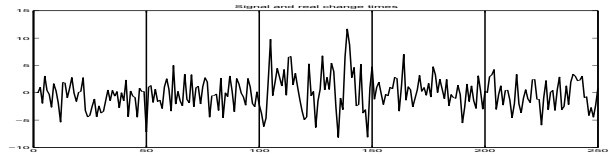


Fig. 7. The signal and the real change instants

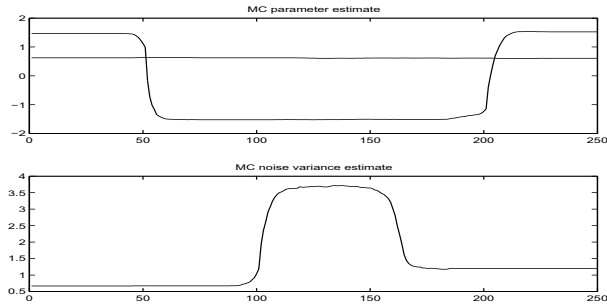


Fig. 5. Model parameter and noise variance estimates obtained in Monte Carlo simulation

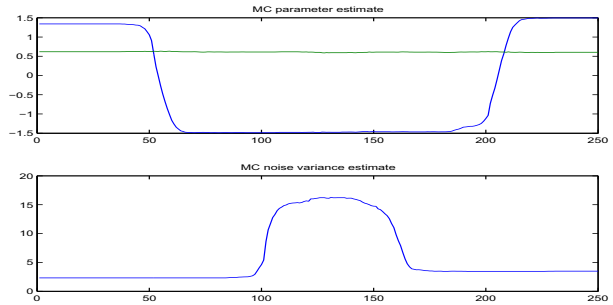


Fig. 8. Model parameter and noise variance estimates obtained in Monte Carlo simulation

Fig. 6.

It can be noted in this case better estimation results for noise variance estimate than in previous case, due to the significant values of noise variance. Due to the increased value of the noise/signal ratio, the histogram from Fig. 6 points out an increase of the performance concerning the real change instants, for noise variance, than for the model parameters.

The last experiment was performed for the model parameters and noise variance values given in Table 3.

The signal resulted by simulation for the same FIR process, a noise realization of the random sequence, e_t , and the real change instants are given in Fig. 7. Fig. 8 presents the model

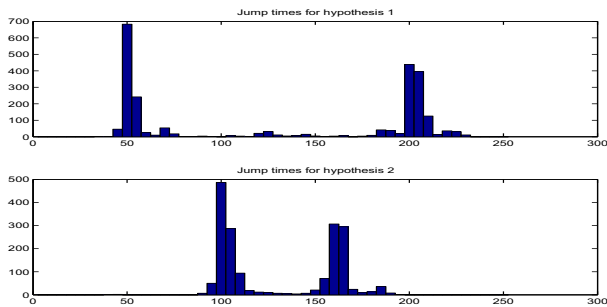


Fig. 6. The histogram for change detection instants in model parameters and noise variance

TABLE III
MODEL PARAMETERS AND λ VALUES

Time	1-50	51-100	101-150	151-200	201-250
θ_1	1.6	-1.6	-1.6	-1.6	1.6
θ_2	0.64	0.64	0.64	0.64	0.64
λ	2.	2.	20.	2.	2.

parameter and noise variance estimates obtained in Monte Carlo simulation for 1000 noise realization. The resulted histogram for change detection instants is given in Fig. 9.

It can be noted in this case better estimation results for noise variance estimate than in previous cases. The value of λ tends to the real values, while the model parameter estimates are more affected than in the previous case. This is due to the new experiment conditions. The same effect can be noted in Fig. 9 for the histogram of the changes in parameter and noise variance estimates. Also, due to the new increased value of the noise/signal ratio, an increase of the variance of the detected change instants can be noted for model parameters, as well as for the noise variance.

All the presented results point out the effectiveness of the proposed approach, able to discriminate between the changes in model parameters and noise variance changes.

B. Seismic signal results

The procedure was applied to strong motion records obtained in a 12-storey reinforced concrete building, during the

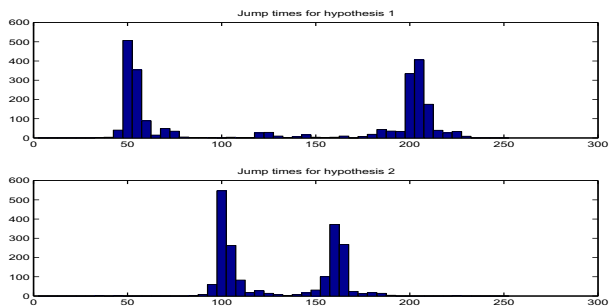


Fig. 9. The histogram for change detection instants in model parameters and noise variance

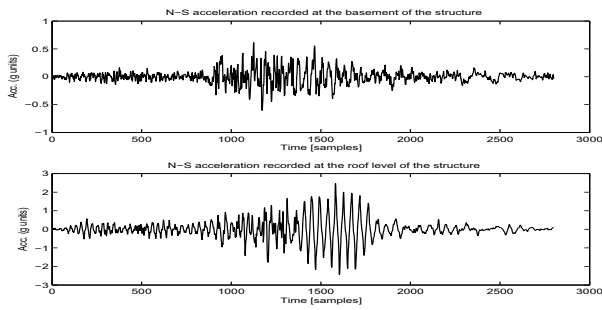


Fig. 10. Input-output data: N-S direction

August 30/31, 1986 Romanian earthquake. The transversal (N-S), longitudinal (E-W) and vertical (VE) components of the acceleration recorded at the basement and at the roof level of the structure have been analyzed. The seismic acceleration and the response acceleration for all components were sampled at 0.02 seconds. We present here only the results for transversal N-S component, represented in Fig. 10.

For regular symmetric structures, such as the investigated building, the identification results obtained assuming a multi-input, single-output system are practically similar to those obtained for a multi-input, multi-output system. The identification results obtained in this case, [7], point out the fact that the transfer function corresponding to the input in the direction of the output is more dominant, suggesting that the building could have been identified using only the input in the direction of the output. This is not surprising, since the structure is symmetric and there is very little torsion.

In this case, the change detection procedure has been applied for 3 single-input single-output systems, corresponding to the investigated directions of seismic wave propagation: N-S, E-W and VE.

The Fourier amplitude spectra of the records point out that their frequency content is mainly in the range 0-5 Hz for transversal and longitudinal components and in the range 0-12 Hz for vertical component. Therefore, the transversal and longitudinal data were low-pass filtered using a zero-phase, second-order Butterworth filter, with the cut-off frequency of 5 Hz; the vertical components were low-pass filtered with a similar filter with cut-off of 12 Hz.

After a preliminary analysis of the seismic components, for different length of the sliding window, we find the length of 200 data (4 sec.) as the best for this application and a model of the following form:

$$y_t = \theta_1 * u_{t-1} + \theta_2 * u_{t-2} + \lambda * e_t \quad (15)$$

Using the procedure, described in the previous section, we present in Fig. 11 the evolution of the parameter estimates and of the noise variance, during the seismic motion on N-S direction, respectively. It can be noted a correspondence between these results and the evolution of the original seismic signals.

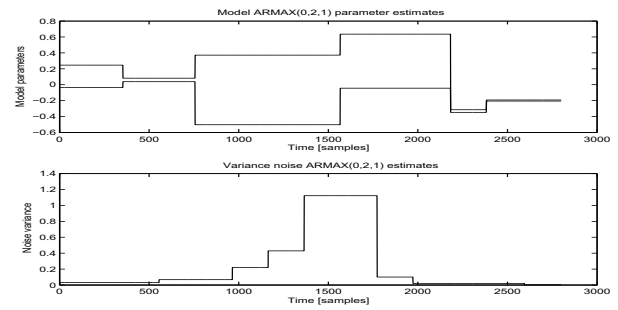


Fig. 11. Model parameter and noise variance estimates: N-S direction

The models obtained at this phase, for each segment, can be used to estimate the modal characteristics of the structure for each direction.

The same seismic data have been used in a change detection procedure presented in [8]. It can be noted (see [8]) that the number of the change instants and their locations, obtained by the both methods, are similar. This confirms the effectiveness of the proposed approach, which is more efficient than the method presented in [8], concerning the computational effort.

IV. CONCLUSIONS

The conceptual description of a new procedure able to discriminate the model parameter and noise variance changes is developed. It is assumed that the data can be described by one linear regression model within each segment with distinct parameter vector and noise variance. It can be used to separate the changes in the experimental conditions from the real changes in a system. Some experimental results obtained by Monte Carlo simulation, for a second order FIR model, and in analysis of seismic signals are presented, proving the effectiveness of the proposed approach.

REFERENCES

- [1] M. Basseville, M., A. Benveniste, G. Moustakides and A. Rougee, "Detection and diagnosis of changes in eigenstructure of nonstationary multivariable systems", *Automatica*, pp.479-489, 1987.
- [2] Th. D. Popescu, "Some experiences with change detection in dynamical systems", *Proc. of the 8th International Conference, KES 2004*, Wellington, New Zealand, pp. 1220-1227, 2004.
- [3] B. Wahlberg, "Robust Frequency Domain Fault Detection/Diagnosis", *Technical Report EE/CIC8902*, University of Newcastle, 1989.
- [4] Th. D. Popescu, and S. Demetriu, "Robust change detection in dynamic systems with model uncertainties", *Prepr. of the 4th IFAC Conference on System Structure and Control*, Bucharest, Romania, pp. 290-296, 1997.
- [5] Th. D. Popescu, "Detection and diagnosis of model parameter and noise variance changes with application in seismic signal processing", *Mechanical Systems and Signal Processing*, pp. 1598-1616, 2011.
- [6] F. Gustafsson, *Adaptive Filtering and Change Detection*, Willey, NJ, 2001.
- [7] Th. D. Popescu, and S. Demetriu, "Identification of a multi-story structure from earthquake records", *Prepr. of the 10th Symposium on System Identification, SYSID'94*, Copenhagen, Denmark, pp. 411-416, 1994.
- [8] Th. D. Popescu, "Change detection procedure with application in structures monitoring subject to seismic motions", *International Journal of Innovative Computing, Information and Control*, pp. 1285-1294, 2009.

With attackers wearing many hats, Prevent your “Identity Theft”

Gundeep Singh Bindra
Computer Science Department
SRM University
Delhi, India
mailbox@gundeepbindra.com

Dhrupad Shrivastava
Computer Science Department
Birla Institute of Technology, Mesra
Rajasthan, India
dhru.2882@gmail.com

Richa Seth
Infosys Technologies Limited
Maharashtra, India
richa_seth@infosys.com

Abstract—An Identity Is Stolen Every 3 Seconds!! Spoofing is a huge issue for most Internet users – in fact, Identity Theft victimized 10.1 Million Americans last year. In this modern digital world, it is very simple to fool the system and pretend to be anyone. With the identity theft getting simpler day-by-day, make sure you are not being digitally spoofed. These days, it doesn't require a computer genius to intrude your privacy or impersonate the victim. And despite various laws, individuals and businesses are still feeling the negative effects of spam alike. The attacker could use your email to send emails to your contacts/anyone or use your caller ID number to call and eventually talk to people. This paper is an attempt to make the readers aware of the various types of spoofing that exists and also how not to fall in prey to them. It is an examination of the various spoofing techniques used by the attackers and their motives. This work intends to increase the awareness and understanding of the Identity Thefts that are and related frauds throughout the world. A few control measures at the users' level are also suggested. This conceptual paper is definitely expected to contribute to future research on similar and related topics as spin off from this study.

Index Terms—Identity Thefts, Spoofing, Email Forging, Call Forging, countermeasures.

I. INTRODUCTION

“Identity theft is a crime. Identity theft and identity fraud are terms used to refer to all types of crime in which someone wrongfully obtains and uses another person's personal data in some way that involves fraud or deception, typically for economic gain.” as defined by U.S. Department of Justice. [1] Identity thieves steal key pieces of personal information and use it to impersonate you and commit crimes in your name. If you are a victim, you could end up spending many hours trying to clear your name and may suffer emotional anguish throughout the process. In extreme cases, you could also suffer a loss of reputation, as court judgements for bad debts could be registered against you and your credit rating could tumble. This, in turn, could make it difficult for you to find employment or get access to credit when you need it. [2] The Federal Trade Commission reported that 13 million Americans suffered from identity theft in 2010, and stated that “ID theft costs consumers about \$50 billion annually” (Finklea,2010) [3] In spite of the attempts to enforce the law, the number of new identity (ID) theft victims is increasing everyday across the

globe. ID theft “can refer to the preliminary steps of collecting, possessing, and trafficking in identity information for the purpose of eventual use in existing crimes such as personation, fraud, or misuse of debit card or credit card data” (justice.ca, 2010). Any factual or subjective information recorded or not, about an identifiable individual is personal information. This includes such things as your name, address, age, gender, identification numbers, credit card numbers, income, employment, assets, liabilities, payment records, personal references and health records. It can also include information about your purchasing preferences, family (such as mother's maiden name), interests, or attitudes. [4]

II. STATISTICS

Approximately 15 million United States residents have their identities used fraudulently each year with financial losses totaling upwards of \$50 billion. On a case-by-case basis, that means approximately 7% of all adults have their identities misused with each instance resulting in approximately \$3,500 in losses. Close to 100 million additional Americans have their personal identifying information placed at risk of identity theft each year when records maintained in government and corporate databases are lost or stolen. These alarming statistics demonstrate identity theft may be the most frequent, costly and pervasive crime in the United States. [5]

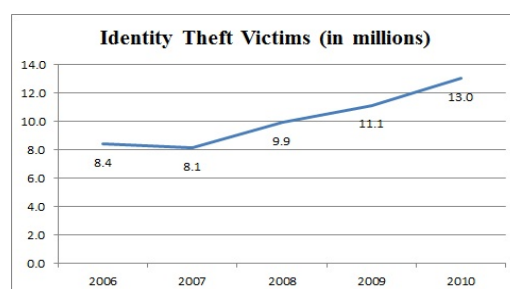


Fig. 1. Identity Theft Victims [6]

President Obama introduced the National Strategy for Trusted Identities in Cyberspace Plan in April 2011. The new plan calls for a proposed single authentication system that acts as a sole online identity for consumers. The plan, however, may inversely increase the extent of fraud committed against Americans. Essentially creating a master key to ones online

identity can have extensive ramifications if the single key is lost. Over the past 2 years, an estimated total of 11.7 million Americans were victims of some sort of identity theft. Over a 5-year period, this number is increasing at an average rate of 1.1 million Americans, representing 10.1 million victims per year. [6] (Statistics according to the FTC).

A. Identity Fraud Statistics For 2012

- ≡ Over 300 identity fraud related arrests per year.
- ≡ Nearly 12 million people are affected by identity fraud each year and nearly 3 million of those identities are that of deceased people (yes, dead!).
- ≡ Research conducted by Javelin Strategy & Research say that although the amount of money stolen due to identity fraud has remained steady, identity fraud itself has increased by more than 13% in the last year alone. Also, people with social media files such as Facebook put themselves at a higher risk for identity fraud due to the personal information they put on their profiles.
- ≡ The research also shows that people with smart phones, like the iPhone and Androids, are at a higher risk for identity fraud. A survey showed that more than 7% of smartphone owners were victims of identity theft, that's 1/3 more than the general public. If you are a smartphone owner and are worried about being a victim of identity theft there are two simple things you can do. Firstly, install the available OS updates when you are prompted or they become available. Keep your phone up to date! And secondly, password protect your phone and to be even safer make sure to change your password at the most every 72 days. [7]

III. IP (INTERNET PROTOCOL) SPOOFING

In computer networking, the term IP address spoofing or IP spoofing refers to the creation of Internet Protocol (IP) packets with a forged source IP address, called spoofing, with the purpose of concealing the identity of the sender or impersonating another computing system. [8].

In a spoofing attack, the intruder sends messages to a computer indicating that the message has come from a trusted system. To be successful, the intruder must first determine the IP address of a trusted system, and then modify the packet headers to that it appears that the packets are coming from the trusted system. In essence, the attacker is fooling (spoofing) the distant computer into believing that it's a legitimate member of the network. The goal of the attack is to establish a connection that will allow the attacker to gain root access to the host, allowing the creation of a backdoor entry path into the target system. [9]

There are many proxy lists available on the Internet that lists the various IPs that can be manually configured in the browser resulting in a successful IP Spoofing method. Other simpler methods is application-based proxy, where the software is installed and it lets you navigate the Internet with an extra layer of privacy and security on any web enabled platform. Examples: Web proxies provide a quick and easy way to change your IP address while surfing the Internet. Web proxies

are extremely portable, as they do not require the installation of additional software or modification to computer networking settings. They are used like a search engine, except that you enter a website address instead of a search query into a form, and web proxies return webpages rather than search results. The sites you visit through the proxy see an IP address belonging to the proxy rather than your IP address. Examples: hidemyass.com; flyproxy.com; proxify.com; proxyapp.org etc.

IV. MAC (MEDIA ACCESS CONTROL) SPOOFING

MAC spoofing is a technique for changing a factory-assigned Media Access Control (MAC) address of a network interface on a networked device. In 2011, Aaron Swartz was charged under the Computer Fraud and Abuse Act; part of the indictment against him related to his alleged use of MAC spoofing to download articles from JSTOR using the MIT network. [10]. To change your MAC address in Windows, go to Device Manager -> properties -> double click on the NIC.

- ≡ In the General tab, click on the Configure button.
- ≡ Click on Advanced tab. In the Property section, select Network Address or Locally Administered Address.
- ≡ To the right, "Not Present" radio button is by default selected as value. Change the value by clicking on radio button for Value:, and then type in a new MAC address to assign to the NIC. Click OK when done.

To change your MAC address in Linux (and most *nix system), all it takes is two easy to script commands:

- ≡ `ifconfig eth0 down hw ether 00:00:00:00:00:01`
- ≡ `ifconfig eth0 up`

The shell command to change the MAC address in Sun Solaris is as below:

- ≡ `sudo ifconfig en0 ether aa:bb:cc:dd:ee:ff or`
- ≡ `sudo ifconfig en0 lladdr aa:bb:cc:dd:ee:ff (Mac OS X)`
- ≡ `ifconfig <interface> <class> <address> (Sun Solaris)`

where en0 is the network interface (numbered from en0, en1, en2 ...) and aa:bb:cc:dd:ee:ff is the desired MAC address in hex notation.

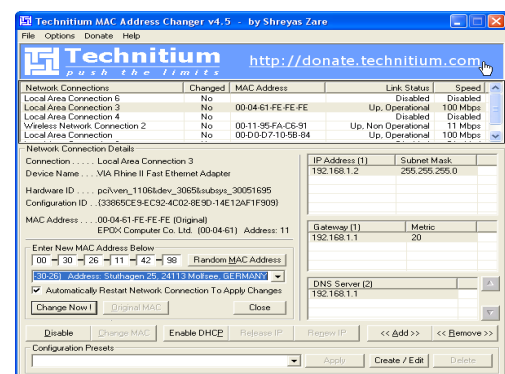


Fig. 2. Technitium MAC Address Changer – Default Screen

Alternative: Use GNU Mac Changer for Linux and other *nix systems, Mac Daddy for MAC OS and Macshift, SMAC by KLC Consultings or my favorite freeware Technitium MAC Address Changer v6 for Windows.

V. EMAIL SPOOFING

Email spoofing is email activity in which the sender address and other parts of the email header are altered to appear as though the email originated from a different source. Because core SMTP doesn't provide any authentication, it is easy to impersonate and forge emails. Although there are legitimate uses, these techniques are also commonly used in spam and phishing emails to hide the origin of the email message. By altering an email's identifying fields, such as the From, Return-Path and Reply-To (which can be found in the message header), email can be made to appear to be from someone other than the actual sender. Authenticated email provides a mechanism for ensuring that messages are from whom they appear to be, as well as ensuring that the message has not been altered in transit. Similarly, sites may wish to consider enabling SSL/TLS in their mail transfer software. Using certificates in this manner increases the amount of authentication performed when sending mail.

Services like MailFool (mailfool.com) allows you to enter any e-mail address as the sender of an e-mail, and it is very difficult to distinguish it from a genuine e-mail from that person! You simply say who you want to send the e-mail to, who you want the e-mail to appear to be from, and your real name and e-mail address. you will then be e-mailed a confirmation password which you must enter on the confirm page on this site, and the e-mail will be sent! But through the message ID mentioned at the end of the email, the recipient can track the original sender of the mail and the sender can be notified of this act. There is another category of service where on the homepage they mention a warning message but not in the email. The warning message mentioned on sendanonymousemail.net is something like this "SendAnonymousMail is not liable for your emails you send at any time. Your IP Address is x.x.x.x so don't do anything illegal. If you send death threats, abuse, slander or anything illegal we WILL publish your IP address and block you from this site" Another example is Anonymailer.net anonymailer.net/send-free-email.asp

Then there are many others which do provide you the facility to send an email being anyone you want but for safety reasons they mention it in the mail itself. Fogmo [18] (previously hoaxmail.com) displays a message "The following message was sent using Fogmo:" but it can be removed by buying credits with money. At the time of testing this was detected by Gmail Spam displaying this message "This message may not have been sent by..." warning". Deadfake.com [19] also says "(No, this email's not real, it's <http://deadfake.com>)" before the actual mail, but went undetected at the time of testing. Also in this category comes fakesend.com. The domain Funmaza.co.uk (mailz.funmaza.co.uk) goes a step ahead, it doesn't mention any warning message either on the email or on the homepage and what makes it dangerous that it goes undetected by Gmail (at the time of writing this September 4, 2012).

If that did raise red alert or wasn't that dangerous for you let me show you the best so far online application for Email spoofing - Emkei's Fake Mailer [11].



Fig. 3. Emkei's Fake Mailer [11]

This is the most sophisticated option as it has a numerous of features besides going unnoticed by the Gmail Spam filters. This service also lets you format your text, add attachments, change the default reply-to address etc. Eventually every website does maintain a record of the IP address of the users visiting the website but that can be easily taken care of. (Refer Section III: IP Spoofing)

TABLE I. FAKE EMAIL SERVICES COMPARISONS

MailServer	Anonymity	Free	Captcha	Undetected by Filters	Allows Attachments	Reply-to
<i>MailFool</i>	x	/	x	x	x	x
<i>Sendanonymousemail</i>	x	x	x	x	x	x
<i>Anonymailer</i>	x	x	/	x	x	x
<i>Fogmo</i>	x	x	/	x	x	x
<i>Deadfake</i>	x	x	/	/	x	x
<i>FakeSend</i>	x	x	x	x	x	x
<i>Funmaza</i>	/	/	/	/	x	x
<i>Emkei</i>	/	/	/	/	/	/

Comparisons of different Fake Email Services

VI. CALLER ID SPOOFING

If you have caller ID, you probably assume whatever shows up on the display is accurate and reliable. That could be a big mistake—and a costly one. [12]. First, it is important to understand that the public switched telephone network is really comprised of two distinct networks - a voice network that carries the actual conversation, and a packet data network that controls call setup, tear-down and all of the data necessary for network features like Caller ID, Calling Name and 800 services. This data network is known as the SS7 or "signaling system 7" network. Originally, access to the SS7 network was limited to exchange carrier switches and databases (i.e., calling name database and 800 database) but was extended to PBXs in order to offer Caller-ID and other advanced signaling services to enterprise customers via an ISDN PRI trunk. The advent of voice-over-IP networking and the Session Initiated Protocol (SIP) has extended the visibility of the signaling information to any VoIP device with a proper connection to the network. [13].

A. History

Many people do not realize that Caller ID spoofing has been around since Caller ID was created. For over a decade

Caller ID spoofing was used mainly by businesses with access to expensive PRI (Primary Rate Interface) telephone lines provided by local telephone carriers. In the early 2000's phone hackers, also known as "phone phreaks" or "phreaks", began using Orange boxing to attempt to spoof Caller ID. In late 2003 and early 2004 the same phone phreaks began to explore a relatively new platform for developing voice applications, known as VoiceXML or VXML, which was offered by companies such as Voxeo. In 2005 a handful of new sites allowing you to spoof your Caller ID were quietly launched. Some of the sites were PiPhone.com, CallNotes.net, SecretCalls.net, StayUnknown.com, SpoofTech.com, SpoofTel.com, and SpoofCard.com. Everything seemed to be going smoothly for the Caller ID spoofing industry, but then in late February 2006, SpoofCard and Telespoof both received letters from the FCC notifying them of investigations into their services. As spoofing seems to be getting closer and closer to being regulated by the US government, the Caller ID spoofing industry seems to have slowed down and the only new site that has appeared in 2007 was SpoofEm.com, a white-label version of SpoofTel.com. [14]

B. How it works?

Caller ID spoofing, where the caller manipulates the information that shows up on caller ID, making it seem like they are calling from anywhere they choose, is increasingly common. Scammers use it to trick victims into handing over their money or personal information.

There are different techniques that you can use to do so, when placing a call, the originating switch will populate caller information in two distinct data packets that traverse the SS7 network. The first, and most critical, is the automatic number identification, or ANI. ANI identifies the true billing telephone number of the originating line, is not usually displayed to a called line and typically cannot be manipulated by an end user as it is populated by the serving CO switch. However, in the VoIP world it is possible to originate a call that has a user-defined ANI, which may be arbitrary. ANI information is sent to 911 centers (PSAPs) and can be delivered to PBXs over an ISDN PRI. The delivery of ANI to a called party is typically an option on an inbound 800 line to a call center operation, but it may also be visible to an IP device connected via a SIP trunk. The second packet of number information sent with every call is the Caller ID. This is the number that will be displayed to the called party when the call arrives at the destination. The Caller ID (CID) can be blocked, either by invoking a feature code when the call is originated (per call blocking) or by subscribing to a privacy feature (per line blocking). However, neither feature actually removes the CID information from the call, but instead sets a privacy signaling flag that indicates to the terminating office that the CID is not to be delivered to the called party. CID can be manipulated, changed, or spoofed by a switch that is connected to the signaling network. This can be a central office switch, a PBX with an ISDN PRI trunk, or a VoIP switch with a SIP trunk connection. The most used PBX open source software is Asterisk. Asterisk powers the IP PBX systems, VoIP gateways, conference servers and more. There is

a function in Asterisk that allows you to set the Caller ID number, "Set(CALLERID())". This makes it easier for programmers because it gives you more flexibility with your calls and a direct connection with the VoIP provider. Another technique is by writing PERL or PHP scripts that allow you to do more than just spoof the phone call. These scripts can allow the caller to change their voice, record the phone call and have it emailed to the caller.

There are many other simpler methods that are used to carry out the same; caller ID is a favorite tool. Caller ID spoofing doesn't require a computer genius. In fact, it's easier than you might think. There are lots of web sites that sell spoofing "calling cards" which make spoofing as simple as just punching in some numbers. Other sites enable spoofing via a web-based system. "It's as easy as making a phone call," says Robert Siciliano, CEO of IDTheftSecurity.com. Most caller ID spoofing services only require a credit card to sign up and don't care or police how the service is used.. Many Caller ID spoofing service providers also allow customers to initiate spoofed calls from a Web-based interface. Some providers allow entering the name to display along with the spoofed Caller ID number, but in most parts of the United States, for example, whatever name the local phone company has associated with the spoofed Caller ID number is the name that shows up on the Caller ID display. SpoofCard[25a] offers you the ability to change or spoof what someone sees on their caller ID display when they receive a phone call. Crazy Call [15] is the ultimate tool for making prank calls and fooling your friends. You can also change the pitch of your voice for deep and creepy or high and funny. The site is up [http://downforeveryoneorjustme.com/crazycall.net] but everywhere where the laws are restricted the website doesn't open. This issue can be easily be resolved by using a web proxy such as Hidemyass.com or flyproxy.com.



Fig. 4. CrazyCall – Caller ID Spoofing Tool [15]

The app is not working in the USA because of the Truth in Caller ID Act. The app is available for download for iPhone and android phones making it easier to make spoof calls.

C. Legal Issues

If you thinking that spoofing caller ID is always used with a bad intent or to make prank calls. Hold on ! there are many valid reasons to spoof caller ID as well. [16]. Lawmakers have made several attempts to address this issue, though so far without any success. The latest attempt is a bill sponsored by Sen. Bill Nelson of Florida. Clearly the technology exists to

create a spoofed Caller ID. However, the FCC does have an opinion on the blocking or spoofing of CID by telemarketers and addresses this issue very specifically at [17]. The Truth in Caller ID Act of 2009, which was signed into law Dec. 22, 2010, prohibits caller ID spoofing for the purposes of defrauding or otherwise causing harm. In June 2010, The Federal Communications Commission (FCC) adopted rules implementing the Truth in Caller ID Act.

1) FCC Rules:

- ≡ Prohibit any entity for transmitting misleading /inaccurate CID information with the intent to defraud, cause harm, or wrongfully obtain anything of value.
- ≡ Subject violators to a penalty of up to \$10,000 for each violation of the rules.
- ≡ Exempt authorized activities by law enforcement agencies and situations where courts have authorized caller ID manipulation to occur. [18]

D. Anti-Caller ID Spoofing:

The only research that has been done with this issue is from a company called TelTech, the owner of the application TrapCall, which allows you to view who's calling you if the caller's number is blocked/private. However, their application does nothing about spoofed phone calls. There is an attempted to create an application to prevent Caller ID spoofing. Two different approaches to are suggested. First, was to contact the cell tower and request the phone number that was calling the recipient's cellular phone. A cellular phone sends a message to the cell tower every few seconds to alert the cell tower that the cellular phone is available to place and receive calls and/or messages. This allows the tower to know the Caller ID number before it is changed. The process of doing this is simply writing a request to the cell tower and switch to find the caller of the incoming call. The second approach was to call back the phone number of the incoming phone call. By calling back the phone number I am looking for a tone stating that the phone is busy. If the tone is busy it will tell that there is a possibility that this is truly the incoming call. The process of doing this is to place the incoming call on hold and be able to dial out the number to the incoming call, using the already created Android source code modify a few of the telephony methods. In the first approach there is a higher percentage of getting the correct CID. [19]

CONCLUSION

As explicitly illustrated in the paper that an attacker can easily spoof and IP Address, MAC Address, an email ID and even Caller ID. This raises an alarm that with such powers of using someone else's email to send out an email and using someone else's phone number to make a call, even worse with a spoofed IP and MAC makes it virtually impossible to catch the attacker. But here we focus on making our readers aware that such things do exist not simply take what the CID or Sender's Email ID for granted. Scammers will say anything to get you to divulge data. In case of Call Spoofing, if you feel you are being swindled, to hang up and call back. But don't call back the number they give you. Call a number that you get online or in the phonebook.

REFERENCES

- [1] Bureau of Justice Assistance, National Crime Prevention Council, "Preventing Identity Theft: a Guide for consumers". Page 2, "What's Identity Theft". <http://www.ncpc.org/cms-upload/prevent/files/IDtheftrev.pdf>
- [2] "Identity Theft: Are you a victim?". ISBN 978-0-662-44825-9. [http://cmcweb.ca/eic/site/cmc-cmc.nsf/vwapj/Are%20you%20a%20Victim.pdf/\\$FILE/Are%20you%20a%20Victim.pdf](http://cmcweb.ca/eic/site/cmc-cmc.nsf/vwapj/Are%20you%20a%20Victim.pdf/$FILE/Are%20you%20a%20Victim.pdf)
- [3] Ali Hedayati, "An analysis of identity theft: Motives, related frauds, techniques and prevention". DOI: 10.5897/JLCR11.044. ISSN 2006-9804©2012 . <http://www.academicjournals.org/jlcr/PDF/pdf%202012/Jan/Hedayati.pdf>
- [4] Credit Counseling Services of Atlantic Canada, Inc, "Empowering for the future Identity Theft" http://www.solveyourdebts.com/pdfs/identity_theft.pdf
- [5] identitytheft.info, "Identity Theft Victim Statistics". <http://www.identitytheft.info/victims.aspx>
- [6] John Borkowski, "Ken Wisnefski Discusses the Problem with NSTIC Plan". September 23, 2011. <http://www.webimax.com/blog/ken-wisnefski/ken-wisnefski-discusses-the-problem-with-nstic-plan>
- [7] Hubpages, "Identity Theft Statistics 2012". <http://gpluspro.hubpages.com/hub/Identity-Theft-Statistics-2012>
- [8] Tanase, Matthew (March 11, 2003). "IP Spoofing: An Introduction". The Security Blog. http://66.14.166.45/sf_whitepapers/tcpip/IP%20Spoofing%20-%20An%20Introduction.pdf Retrieved February 10, 2012.
- [9] Victor Velasco, "Introduction to IP Spoofing" SANS Institute InfoSec Reading Room. November 21, 2000. http://www.sans.org/reading_room/whitepapers/threats/introduction-ip-spoofing_959
- [10] Internet Activist Charged in M.I.T. Data Theft By NICK BILTON July 19, 2011, 12:54 PM. <http://bits.blogs.nytimes.com/2011/07/19/reddit-co-founder-charged-with-data-theft/>
- [11] Emkei, "Online fake mailer with attachments, HTML editor and advanced settings...". <http://emkei.cz>
- [12] Fiber Net Monticello, "CALLER ID SPOOFING TIPS" http://www.monticellofiber.com/PDF/FNM_Caller_ID_Spoofing_Document.pdf
- [13] Abilita, Independent Communication Expertise, "SPOOFING CALLER ID". <http://www.abilita.com/aweigand/Articles/Caller%20ID%20Spoofing.pdf>
- [14] calleridspoofing.info, "What is Caller ID Spoofing?". <http://www.victimsofcrime.org/docs/src/what-is-caller-id-spoofing.pdf?sfvrsn=2>
- [15] Crazy Call, "CrazyCall is the ultimate tool for making prank calls and fooling your friends.". <http://www.crazycall.net>
- [16] Wikipedia, "Caller ID Spoofing". Section: Valid reasons to spoof caller ID . http://en.wikipedia.org/wiki/Caller_ID_spoofing
- [17] Federal Communications Commission, "Caller ID and Spoofing". <http://www.fcc.gov/cib/consumerfacts/callerid.html>
- [18] FCC's Consumer & Governmental Affairs Bureau, "Caller ID and Spoofing" <http://transition.fcc.gov/cgb/consumerfacts/callerid.pdf>
- [19] Janette Archie, "Reversing Caller ID Spoofing". <http://www.cse.sc.edu/files/Janette%20Archie.pdf>

Fully distributed certificate authority based on polynomial over elliptic curve for MANET

Ahmad Alomari

Faculty of Mathematics and Computer Science, University of Bucharest, Bucharest, Romania
alomari.jordan@gmail.com

Abstract— A mobile ad hoc network (MANET) is a wireless communication network, which does not rely on any centralized management or a pre-existing infrastructure. Various certificate authorities (CAs) distributed over the network, each with a periodically updated share of the secret key, is usually adopted. Elliptic Curve Cryptography (ECC) is a cryptographic technique prominent suited for small devices, like those used in wireless communications, and is gaining momentum. The main advantage of ECC versus RSA is that for the same level of security it requires a much shorter key length. The purpose of this work is to design and implement a fully distributed certificate authority based on polynomial over elliptic curve, and based on trust graphs and threshold cryptography. Which though has better cryptography in nature and that their security is based on the elliptic curve discrete logarithm problem difficult solution.

Keywords- Certificate Authority (CA); MANET; ECCRL (certificate revocation list)

I. INTRODUCTION

A Mobile Ad hoc Network (MANET) is a system of wireless mobile nodes that can communicate with each other without the use of predefined infrastructure or centralized administration. People and vehicles can thus be Internet worked in areas without a preexisting communication infrastructure or when the use of such infrastructure requires wireless extension. In the mobile ad hoc network, nodes can directly communicate with each other within their radio ranges; if the nodes that not in the direct communication range they use the intermediate nodes to communicate with each other. In these two situations, the network forms when all the nodes participated in the communication automatically, therefore this kind of wireless network can be known as mobile ad hoc network. Mobile ad hoc networks (MANETs) have become one of the fastest growing areas for the researchers, with the propagation of cheaper, smaller, and more powerful mobile devices. Due to self-organize and rapidly deployment, MANET can be applied to different applications including emergency relief scenarios, battlefield communications, public meeting, law enforcement, also the ad hoc self-organization makes them suitable for virtual conferences, where setting up a traditional network infrastructure is a time-consuming high-cost task. Among all the research issues of ad hoc network, the nature of communication and lack of infrastructure support make the security is particularly more challenging. A number of security mechanisms has been developed and proposed, but

still it is still hard to ensure that whole network is free from any malicious attack.

Providing security services including authentication, integrity confidentiality, availability, and anonymity to the mobile user is the main goal of the security solutions for MANET. To achieve to this goal, the security solution should supply complete protection spanning the entire protocol stack. The traditional Internet style key distribution protocols, like Kerberos, relying on online trusted third parties (TTP) to distribute session keys to nodes are infeasible for ad hoc networks because the TTP may be out of range or not available to all of the nodes or during certain times for a number of reasons. These include communication range limitations, network dynamics, node movements and unknown network topology prior to deployment. There exist a number of researches and proposals for ad hoc networks, which try to increase the availability of the key distribution service by repetition the online key server to a subset of nodes arranged arbitrarily or hierarchically. Nevertheless, the performances of these schemes, in terms of efficiency and scalability, are still not reassuring. In additional to, these schemes still need TTP; compromising the TTP compromises all the keys it issues.

This work focuses on peer-to-peer key management in fully Self-organized mobile ad hoc networks. A fully self-organized MANET means any user with the appropriate equipment (and software) can join and leave at random, and we can call this network is an “open” network; there is no form of access control. Such a network will therefore not find application in, for example, hostile military environments, but rather in commercial, community-based environments. Present approaches for authentication services depend on centralized management approaches by either certificate authorities (CA) or key distribution centers. A centralized approach may be acceptable, in cases where a specific node can be protected and is accessible by other nodes of the network. However, for the wireless ad hoc networks that we visualize for our targeted applications, a centralized approach will suffer from a single-point of service denial and may be unreachable by network nodes requiring CA services. Thus a more robust CA approach must be used. This need for wireless ad hoc networks is presently a very active research area. Providing CA functionality in an ad hoc network is to assign a single node to be the CA is the simplest approach. The success of this scheme depends on that single CA node. Since failure of one node breaks the system, this approach is not fault tolerant. Similarly this approach is highly

vulnerable, since an adversary need only compromise one node to acquire the secret key. Finally, given the unpredictability and expected mobility of ad hoc networks, it may be possible that nodes will not be able to reach the CA In due course, making availability greatly unpredictable. Thus, a single CA cannot effectively service a whole ad hoc network.

In this paper we proposed a dynamic fully distributed certificate authority scheme based on a polynomials over elliptic curve for Mobile Ad Hoc Networks, which though has better cryptography in nature and that their security is based on the elliptic curve discrete logarithm problem difficult solution, but the participants' keys are distributed by a trusty center, which takes a lot of inconvenient in practical applications. This article offered a sharing scheme based on a polynomial over elliptic curve, in these scheme participants would hold optional sub- secret keys.

II. RELATED WORK

One of the first approaches to solve the key management problem in MANETs is Partially Distributed Certificate Authority Approaches published in [1]. The authors Zhou and Z. J. Haas, proposed a distributed public key management service for asynchronous ad hoc networks, where the trust is distributed between a set of nodes by allowing the nodes share the system secret. The distributed certificate authority (DCA), illustrated in Figure (1) [1], consists of n server nodes which, as a whole, have a public/private key pair K/k . The public key K is known to all nodes in the network, whereas the private key k is divided into n shares ($s_1, s_2, s_3, \dots, s_n$), one for each server.

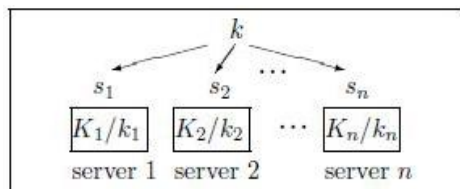


Figure (1): Key management service K/k configuration

The distributed certificate authority (DCA) signs a certificate by producing a threshold group signature as shown in Figure-(2) [1]. Each node generates a partial signature using its private key share and submits the partial signature to a combiner C . The combiner can be any node and requires at least $t + 1$ shares to successfully reconstruct the digital signature.

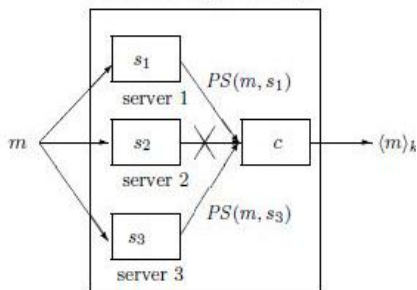


Figure (2): Threshold signature K/k generation

Fully Distributed Certificate Authority Approaches, this solution is first described by Luo and Lu in [2]. It uses a (k, n) threshold scheme to distribute an RSA certificate signing key to all nodes in the network. It also uses verifiable and proactive secret sharing mechanisms to compromise of the certificate signing key and protect against denial of service attacks.

This solution is aimed across planned, long-term ad hoc networks with nodes capable of public key encryption. However, since the service is distributed among all the nodes when they join the network, there is no need to choose or elect any specialized server nodes. Their solution also uses an (n, k) threshold signature scheme to form a distributed certificate authority (DCA). They enhance the availability feature of Practical PKI (public key infrastructure) for Ad Hoc Wireless Networks [3] by choosing n to be all the nodes in the network. The private key SK of the DCA is thus shared among all the nodes in the network and enables a node requiring the service of the DCA to contact any k one-hop neighbor nodes. In contrast to Practical PKI for Ad Hoc Wireless Networks no differentiation is made between server and client nodes with respect to certification services. The solution includes a share update mechanism to prevent more powerful attackers from compromising the certification service.

Implementing a fully distributed certificate authority in an OLSR MANET was proposed D. Dhillon, T. S. Randhawa, M. Wang, L. Lamont [4] they present our approach to integrate a fully distributed CA in a proactive ad hoc routing protocol named OLSR (Optimal Link State Routing). IETF's MANET working group has identified OLSR as one of the four base routing protocols for use in ad hoc networks. The other three are AODV (Ad-hoc On-Demand Distance Vector), DSR (Dynamic Source Routing) and TBRPF (Topology Broadcast Based on Reverse-Path Forwarding) routing protocols. Their approach addresses the concerns of control traffic overload by tightly coupling the operations of a fully distributed CA at the network layer level. The existing OLSR specific packet types, identified in the IETF's draft proposal on OLSR, are used, as much as possible, to also support the proposed PKI. A real test-bed has been constructed in which the existing implementation of OLSRv4 [5] was utilized and a fully distributed CA was introduced. This is to their knowledge the first attempt to address the security issues of OLSR. The paper thus provides valuable insight by detailing the implementation and evaluation of the proposed approach.

III. PRELIMINARIES

As we mentioned earlier, our approach is based on a polynomials over elliptic curve for Mobile Ad Hoc Networks, so, in this section we will review some preliminary concerning these technique Elliptic Curve Cryptography (ECC) [6]. An elliptic curve E over a finite field \mathbb{F}_q consists of all the points consists of all the points $(x; y) \in \mathbb{F}_q \times \mathbb{F}_q$

$$Y^2 + a_1XY + a_3Y = X^3 + a_2X^2 + a_4X + a_6;$$

With $a_i \in \mathbb{F}_q$, whose discriminant is non null, along with the point at infinity. There is a point addition operation

whose neutral element is the point at infinity. This set of points under this operation is an Abelian group. Therefore, a point $Q \in E(\mathbb{F}_q)$ can be multiplied by a scalar:

$$eQ = \underbrace{Q + \dots + Q}_e = P$$

The inverse problem (i.e., given P and Q, find an e such that $P = eQ$), called the Elliptic Curve Discrete Logarithm Problem (ECDLP), discern to be computationally hard to solve. There are several cryptosystems, whose security is based on the intractability of the ECDLP problem. The main concern of the ECDLP, compared to the ordinary DLP for multiplicative groups, is that there exist subexponential algorithms such as the index calculus to solve the DLP on multiplicative groups, but they cannot be used to solve the ECDLP. Hence, it turns to be a harder problem. Under a practical point of view, it appears that shorter keys can be used in the ECDLP while offering the same security as DLP.

IV. DISTRIBUTED CERTIFICATE AUTHORITY BASED ON POLYNOMIAL OVER ELLIPTIC CURVE

We consider an ad hoc wireless network with m mobile nodes. Nodes communicate with each other with the bandwidth constrained, and insecure channel. The m nodes may be dynamically changing as mobile nodes join, leave, or fail over time. Besides, m is not constrained; there may be a large number of networking nodes. The network provides neither logical infrastructure support nor physical [7], [8]. We have the following assumptions. (1) The public key PK for certificate verification is well known to each node in the network (2) Communication between multi-hop communications is considered less reliable compared with one-hop neighboring nodes. (3) Every node has at least k one-hop valid neighboring nodes. (4) To identify misbehaving nodes among its one-hop neighborhood, every node is equipped with some local detection mechanism. Assume that there is a certification authority (CA) an m participant nodes in the mobile ad hoc network (MANET). CA will dispense a secret key to every participant node in the network, the secret key SK_{CA} can be resumed if and only if the number of participants is not less than t. The CA holds a pair of keys (PK_{CA}, SK_{CA}) , PK_{CA} is the public key known by every one, SK_{CA} is the private key with external confidentiality. IN our design we make extensive use of the polynomial secret sharing and fully distributed CA is based on an approaches described by Shamir [1] and Luo and Lu [2] respectively, and we implement our fully distributed over elliptic curve.

A secret, specifically the exponent of the certificate-signing key SK_{CA} , is shared among all nodes in the network according to a random polynomial of order t-1. A coalition of t nodes with t polynomial shares can potentially recover SK_{CA} by Lagrange interpolation, while any coalition up to t-1 nodes yields any information about SK_{CA} .

1. Initialization of proposed scheme:-

In this scheme we choose a secure elliptic curve $E(\mathbb{F}_q)$ over

the finite fields \mathbb{F}_q (q is a prime number) :-

$$y^2 = x^3 + ax + b \quad q > 3$$

Which $a, b \in \mathbb{F}_q$ and satisfy the equation $4a^3 - 27b^2 \neq 0$.

G is a point over the elliptic curve, with a big prime number order n whose binary-length is at least 160 bits. CA will choose a r-power polynomial $g(x)$; Which could be decomposed as follows:-

$$(g(x)) = g_1(x) g_2(x) \dots g_k(x)$$

$g_i(x)$ is r_i -power polynomial which could not be decomposed, the number of polynomials which prime with $g(x)$ over \mathbb{F}_q is :-

$$\phi_n(g(x)) = n^r \prod_{i=1}^k \left(1 - \frac{1}{n^{r_i}}\right)$$

We change the coordinates of G, G is appoint on elliptic curve whose order is large value n, to polynomial formal:- $G = \langle h(x), h'(x) \rangle_{g(x)}$; $G = \langle h(x), h'(x) \rangle_{g(x)}$ means both polynomials $h(x)$, $h'(x)$ will modulo the polynomial $g(x)$. $H(x)$ is a one-way hash function without collision. The coalition of nodes whose response of CA publishes the public parameters $(E, G, g(x), \phi_n(g(x)), H(x))$.

2. A fully distributed CA by using polynomial over elliptic curve

In our proposed of fully distributed CA is passed on an approach described by Lau and Lu in [2]. We apply the CA in mobile ad hoc network over elliptic curve cryptography (ECC) key pair with public key PK_{CA} , private key SK_{CA} , and modulus $g(x)$. First the dealer initialized t nodes and then these t nodes initialized the rest of the network, in the fully distributed, SK_{CA} is distributed by using Shamires secret sharing method by embedded SK_{CA} as the root of polynomial over the Elliptic Curve $E(\mathbb{F}_q)$, the dealer randomly chooses a point r and at-t power polynomial over the elliptic curve $E(\mathbb{F}_q)$, and also the dealer determine the domain parameters $T = (p, a, b, G, n, h)$, which do not be kept secret.

$$F(x) = a_0 + a_1 x^1 + a_2 x^2 + \dots + a_{t-1} x^{t-1}$$

Where $a_0 = SK_{CA}$, $a_k \in [1, \phi_n(g(x))]$, $(k = 0, 1, 2, \dots, t-1)$, $f(0) = SK_{CA}$.

Each shareholder node with a unique non-zero identity receives a share $s_i = f(i) \bmod g(x)$, where $i = 1, 2, \dots, m$.

With knowledge of at least t shares the polynomial can be evaluate by calculating:-

$$F(x) = \sum_{i=1}^t s_i \cdot l_i(x) \bmod n$$

Where $l_i(x)$ is the Lagrange coefficient defines as

$$l_i(x) = \prod_{i=1, j \neq i}^t \frac{x-j}{i-j}$$

$f(x)$ is kept secret, and then the dealer does the following work:

$$\begin{aligned} \text{Order } r p_i &= (x_i, y_i) \bmod n \\ \text{Compute } y_i &= f(x_i) \bmod n \quad (i=1, 2, \dots, m) \\ R &= r G \bmod n \end{aligned}$$

For CA the dealer or coalition nodes in the network publishes the point R and the parameters y_i , in order to check the validity of the secret sharing of the secret key (SK), CA will compute and publish the parameters $H_i = H(r p_i)$, ($i=1, 2, \dots, m$)

Participant nodes have two proceeding to do here, the first each node in the network select private key P_{riv} over the elliptic curve and after that compute the public key $P_{Ki} = P_{riv} \times G \bmod p$, P_{Ki} is a public key of node I for encryption and verification, P_{riv} is a private key of decryption and signing. The nodes use these keys for encryption and decryption of data packets. A challenge and response protocol can be followed to prove the knowledge of the private key P_{riv} and a certificate proves the association. The second the participant nodes in the network compute $p_i = s_i G = \langle h(x), h'(x) \rangle_{g(x)}$, $i=1, 2, \dots, m$, this parameter is used for verification of the partial certificate when the node i signs his certificate by s_i .

A. Self-Initialization

Secure wireless networks used in our targeted applications are comprised of mobile nodes that may result in nodes joining and leaving the network. As we have said previously, the dealer initialized k nodes and then these k nodes initialized the rest of the network, when a new node enters the network and does not have access to a dealer, an alternative method is necessary for this node to join the coalition of nodes able to provide secret shares. This alternative method is necessary to securely provide the node the ability to generate dynamically new secret shares that are compatible with other coalition nodes already in the network. Luo and Lu [2] propose a distributed self-initialization algorithm to address this problem. In particular, they use a coalition of members already in the network. The coalition communicates interactively to generate partial-secret shares that can be combined to generate the secret share for the new node.

The generation of a secret share for a new node that joins the network is constructed by a coalition SK, of t nodes currently in the network. When a new node wants to join the network, a new node broadcasts an initialization request to the neighbor nodes and when all coalition members receive this request they generate the parameter $p_{i+1} = s_{i+1} G \bmod n$ which is encrypted by the public key of the new node. Also they compute two parameters $r p_{i+1}, y_{i+1}$ and open y_{i+1} :

$$\begin{aligned} r p_{i+1} &= (x_{i+1}, y_{i+1}) \bmod n \\ y_{i+1} &= f(x_{i+1}) \bmod n \end{aligned}$$

Each participant communicates privately with every node in S by exchanging secret information. Each node $j \in SK$,

subsequently returns a shuffled version of its secret share s_i to the new node. A shuffled version is used to protect the value of its secret share s_i

$$s'_i \leftarrow \text{shuffle}(s_i, Sk)$$

The value s_i depends on i 's interaction with the other participants in SK and the current size of the network. Once the new node has obtained the shuffled shares it may construct its secret share s_{n+1} by,

$$s_{i+1} \leftarrow \text{unshuffle}_{k,t}(s'_1, \dots, s'_k)$$

B. Certificate Renewal

Since certificates are only valid for a limited time period they need to be renewed before they expire. When a node p has to renew its certificate, p requests a certificate renewal from a coalition of k neighbor nodes. Each node in this coalition checks that the old certificate has not already expired and that it has not been revoked. If it has been revoked, then the nodes ignore the request, otherwise the request is granted, each of these k server nodes generates a partial certificate with a new expiration date and returns it to node p . Node p then combines the k partial certificates to obtain its updated certificate $\text{cert}_{\text{updated}}$. If any of the nodes are compromised they may generate an invalid partial certificate, which they then send to the combiner. The certificate produced by the combiner will then also be invalid. The node will need to update its certificate with the new public key, if the node changes its private and public keys; this is accomplished in a similar way as the renewal

C. Certificate Revocation

Users can revoke any issued certificate to other users in the instance of suspicion in the public key/identity binding. Similarly users can also revoke their own certificate if they know that their private key has been compromised. The certificate revocation mechanism is based on the assumption that all nodes monitor the behavior of their one-hop neighbors and maintain their own certificate revocation lists. If a node discovers that one of its neighbors is misbehaving, it adds its certificate to the CRL (certificate revocation list) and broadcasts an accusation against the node to the neighbor nodes. Any node receiving this broadcast accusation first checks its CRL to verify that the accusation did not originate from a node whose certificate has been revoked. If the accuser's certificate has been revoked the accusation is ignored. Otherwise, the accusation originated from a valid node, the accused node is accepted and the changes are made to the CRL.

V. SECURITY ANALYSES

The CA for this solution requires an organizational/administrative infrastructure to provide the registration and initialization services. The main benefits of this scheme are its availability and that its polynomial is over the elliptic curve.

The security of our scheme depends on the intractability of the Elliptic Curve Discrete Logarithm Problem (ECDLP).

Consider the equation $Q = k p$ where $Q, p \in E(\mathbb{F}_q)$, and $k < p$. It is relatively easy to calculate Q given k and p , but it is relatively hard to determine k given Q and p . This is called the discrete algorithm problem for elliptic curves (ECDLP). This technique makes the certificate authority more robust against the some kinds of attacks.

The security of ECC depends on how difficult it is to determine k given Q and p . This is referred to the elliptic curve logarithm problem. If we make comparison between the RSA and ECC algorithms by comparable key sizes in terms of computational effort for cryptanalysis. Considerably smaller key size can be used for ECC compared RSA. Thus, there is a computational advantage to using ECC with a shorter key length than a comparably secure RSA.

Since all nodes are part of the CA service, it is adequate that a requesting node has t one-hop neighbors for the CA service to be available. The amount of network wide traffic is also limited.

The cost of achieving this availability is a set of rather complex maintenance protocols, e.g. the share initialization and the share update protocols. A larger number of shares are also displayed to compromise since each node has its own share as compared to only the specialized server nodes in the partially distributed solution. The parameter t therefore may need to be chosen larger since an attacker may be able to compromise a larger number of shares between each share update. This in turn affects the availability of the service. The solution must also provide for a synchronization mechanism in the case of network segmentations.

VI. CONCLUSION

In our scheme we proposed a fully distributed certificate authority based on polynomial over elliptic curve, and based on trust graphs and threshold cryptography, this scheme provide a robust and more secure distributed CA over the MANET. Which though has better cryptography in nature and that their security is based on the elliptic curve discrete logarithm problem difficult solution.

References

- [1] L. Zhou and Z. J. Haas, Securing Ad Hoc Networks. IEEE Networks, Volume 13, Issue 6 1999.
- [2] H. Luo and S. Lu, Ubiquitous and Robust Authentication Services for Ad Hoc Wireless Networks., Technical Report 200030, UCLA Computer Science Department 2000[5] J-P. Hubaux, L. Buttyán and S. Capkun.
- [3] S. Yi and R. Kravets, "Practical PKI for Ad Hoc Wireless Networks," Department of Computer Science, University of Illinois, Technical Report UIUCDCS-R-2002-2273, UIIU-ENG-2002-1717, August 2001
- [4] D. Dhillon, T. S. Randhawa, M. Wang, L. Lamont, Implementing a Fully Distributed Certificate Authority in an OLSR MANET, wireless communication and networking conference, 2004. WCNC. 2004 IEEE.
- [5] L. Christensen and G. Hansen, "OLSR Routing Protocol", <http://hipercom.inria.fr/olsr/>, September 2003.
- [6] V. S. Miller, "Use of Elliptic Curves in Cryptography", Proc. CRYPTO'85, Springer-Verlag, New York, pp. 417-426, 1986.
- [7] Y. Dong, A.-F. Sui, S. Yiu, V. O. Li, and L. C. Hui. Providing distributed certificate authority service in cluster-based mobile ad hoc networks. Elsevier, Computer Communications, May 2007.
- [8] Yuan Yangtao, Liu Quan, Li Fen, A Design of Certificate Authority Based on Elliptic Curve Cryptography, 2010 Ninth International Symposium on Distributed Computing and Applications to Business, Engineering and Science

Cut-off Time Calculation for User Session Identification by Reference Length

Jozef Kapusta, Michal Munk, Martin Drlik

Department of Informatics

Faculty of Natural Sciences, Constantine the Philosopher University in Nitra

Nitra, Slovakia

jkapusta@ukf.sk, mmunk@ukf.sk, mdrlik@ukf.sk

Abstract— One of the methods of web log mining is also discovering patterns of behavior of web site visitors. Based on the found users' behavior patterns that are represented by sequence rules, it is possible to modify and improve web site of the organization. Data for the analysis are gained from the web server log file. These anonymous data represent the problem of unique identification of the web site visitor. The paper deals with less commonly used navigation-driven methods of user session identification. These methods assume that the user goes over several navigation pages during her/his visit until she/he finds the content page with required information. The content page is a page where the user spends considerably more time in comparison with navigation pages. The content page is considered to be the end of the session. Searching of the next content page using navigation pages constitutes a new user session. The division of pages into content and navigation pages is based on the calculation of cut-off time C . The verification of exponential distribution of variable that represents the time which user spent on the particular page is coessential. We prepared an experiment with data gained from log file of university web server. We tried to verify, if the time spent on web pages has exponential distribution and we estimated the value of cut-off time. The found results confirm our assumptions that the navigation oriented methods could be used to proper user session identification.

Keywords- *Web Log Mining. Session Identification. Reference Length. Cut-off Time*

I. INTRODUCTION

Log file of the web server is a source of anonymous data about the user. These anonymous data represent also the problem of unique identification of the web site visitor. If we want to analyze the users' behavior on our web site, it is not necessary to know the identity of each visitor, but, it is very important for us to distinguish the web site visitors. The finding of the user behavioral patterns that are represented as sequence rules is based on the data from log file created by the web server. Sequence rules analysis is useful for further modification or optimization of the web site.

II. RELATED WORK

Many web usage mining approaches and methods were surveyed in [1], [2], [3] and [4]. The last comprehensive surveys on web usage mining have been done by Koutri, Avouris and Daskalaki [5] and Kosala and Blockeel [6].

Facca and Lanzi summarize recent developments in web usage mining research in their very interesting papers [7] and [8] of which we can observe the progress in this popular research area.

Identification of user session boundaries is one of the most important processes in the web usage mining for predictive prefetching of user next request based on their navigation behavior. There are a lot of papers which deal with this very actual research problem [9,10,11]. For example, Chitraa and Thanamani [12] presented a novel technique for user session identification in web usage mining preprocessing in their paper. Arumugam [13] described new techniques for identification of user session boundaries by considering IPaddress, browsing agent, intersession and intrasession timeouts, immediate link analysis between referred pages and backward reference analysis without searching the whole tree representing the server pages. He also compared the performance of the given approach with the existing reference length method and maximal reference method.

III. WEB ACCESS LOG SET

Let U be the set of possible requests on web server. This set contains all requests on web server including static web pages, dynamically generated web pages, script files, images etc.

When we clean and remove insignificant requests we obtain the set $WALS$. $WALS$ then represents all accesses to the web stored in log file of web server (Web Access Log Set - $WALS$) and we can define it as:

$$\begin{aligned} WALS \\ = \langle UIP, Date, Method, URI, Version, Status, Bytes, ReferURI, BrowserOS \rangle \end{aligned} \quad (1)$$

$WALS$ is the subset of set U and the whole set $WALS$ is ordered sequentially according to time, i.e. according to item $Date$.

Then we can define session as ordered list of web pages which were visited by one particular user:

$$USS = \langle USID, \langle URI_1, ReferURI_1, Date_1 \rangle, \dots, \langle URI_k, ReferURI_k, Date_k \rangle \rangle, \quad (2)$$

where $1 \leq k \leq n$ and n is record count in *WALS*.

Each session in *USS* is then represented by the set of records stored in log file which have the same *USID* and which are ordered according to time of web page visit.

It is evident that following assertions for the set *USS* are true:

1. For each *USID*, if $1 \leq i \leq j \leq k$ than $Date_i < Date_j$;
2. Each record from *WALS* must belong to some session and each item of *USS* must be a part of *WALS*;
3. Each record from *WALS* must be a part of exactly one user session.

We found also another user session definition $\langle URI_1, \dots, URI_k \rangle$ in [14], where time was neglected, but we do not take it into consideration because we need for the purpose of this paper more complex of *USS*.

Turn our attention now to the problem of path completion among web pages which have been visited by one user and to the corresponding methods. Therefore it is necessary to define a set of web pages visited by one particular user (Path Set - *PS*):

$$PS = \langle USID, \langle URI_1, Date_1, RLength_1 \rangle, \dots, \langle URI_k, Date_k, RLength_k \rangle \rangle \quad (3)$$

where $1 \leq k \leq n$ and *RLength* is time spent on given web page.

Each *USID* is given during the process of path completion, while *RLength* is calculated for each web page. The variable *RLength* is mostly calculated as the difference between timestamp of actual and following web page:

$$RLength_i = Date_{i+1} - Date_i. \quad (4)$$

If we want to use methods for user session identification we have to choose from these two options:

1. All activities realized by one user may be aggregated.
2. All activities which belong to one user session have to be aggregated separately.

We have to bear in mind that the visitor can visit the web site more than one time. It means that the log file can contain multiple records that represent several user sessions.

Individual visitors can be differentiated also based on the identification of sessions. The aim of session identification is to divide individual accesses of each user into separate relations. These relations can be defined in various ways [15].

The reconstruction of individual visitor activity is relatively complicated process. The separation of user session on the basis of IP addresses is the simplest solution. But we must note the fact that IP addresses are not suitable in general for mapping and identification of individual site visitors.

Currently it is not rare that several users share a common IP address, whether they are situated under a certain NAT (Network Address Translation), or proxy equipment. Authentication mechanisms can facilitate identification of the user. However, their usage is undesirable due to privacy protection [16].

The basic session identification method comes from the assumption that the session is represented as a bounded set of clicks realized in defined time.

As we mentioned earlier the main method for session identification assumes that the session is the bounded set of user's clicks realized in defined time interval.

We can use several methods of time based user session identification:

1. We can consider the session to be a set of user's clicks during selected time period, for example during 30 minutes, 10 minutes etc. [16]. It follows the duration of session cannot be greater than θ . It means in the case of set *USS*:

$$Date_k - Date_1 \leq \theta, \quad (5)$$

where $Date_k$ is the last record of the session.

All other records of the log file with timestamp greater than $Date_1 + \theta$ belong to the next user session.

2. The second, more effective, method expects that the session is identified on the basis of sufficiently long interval of time among two recorded visits of the web page. If we denote it as σ than holds that:

$$Date_{i+1} - Date_i \leq \sigma. \quad (6)$$

If this inequality is not true for two consecutive records of log file than these records belong to two different user sessions.

The identification of web site visitors with regard to used web browser belongs to the next method of user identification. This method allows dividing the records from one IP address into several sessions using information about used web browser.

Another method of user session identification uses the parameter referrer in log file and web site map for tracking of user activity on web site. If the log file contains two consecutive records from one IP address and these two web pages are not linked directly this heuristics assumes that these records belong to the two different accesses of two different users from one IP address.

The slightly different method called h-ref heuristic method takes also into consideration time and parameter referrer from log file where it is possible to [16].

Two consecutive records belong to the one session in the selected time window σ , if:

$$ReferURI_i = URI_{i-1} \quad (7)$$

or in other words this equality is not true, $ReferURI_i$ is not defined and :

$$Date_i - Date_{i-1} \leq \sigma. \quad (8)$$

We verified the importance of above mentioned pre-processing methods in several experiments. We described them in detail in [17-19]. We examined different steps of the analysis of the anonymous web site visitors with the aim to recognize the most important one. Observed results are very important from the web site administrator point of view, because they can approve the structure and content of the web site.

Described experiments have been realized since 2009 and they have been focused on the comparison of results of sequence analysis of the log files with different level of data preparation. The examined log files came from the university web site. The aim of the experiments was to examine the necessity of particular steps of data preparation and subsequently their integration and automation.

IV. WEB SITE SEARCHING MODEL

The model of web site searching and model of user behavior are fundamental to correct aggregation of individual user's clicks to meaningful sessions that are sometimes referred to as transactions. We can organize individual web pages of the examined web site to three groups in term of model:

1. content pages,
2. navigation (auxiliary) pages,
3. multiple purpose pages.

The content pages are web pages where the user can find required information. These pages are the reason of visit of individual user throughout his browsing of web space. Therefore we can say that in the case of association rules searching content pages is the most important and our objective is to discover useful rules among those pages.

The other mentioned groups of web pages are necessary for successful site navigation or as sources of auxiliary information (Figure 1).

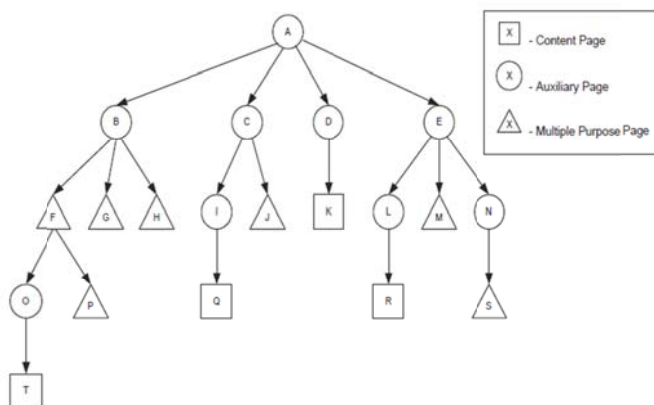


Figure 1. The example of web site map in consideration of different types of web pages [15].

We have to note the fact that the division of web pages into above mentioned three groups may be very individual in term of user model, i.e. particular content page defined by one user can be considered by another user as an auxiliary page [15]

In order to group individual web page references into meaningful transactions for the discovery of patterns such as association rules, an underlying model of the user's browsing behavior is needed. For the purposes of association rule discovery, it is really the content page references that are of interest.

The other page types are just to facilitate the browsing of a user while searching for information, and will be referred to as auxiliary pages. What is merely an auxiliary page for one user may be a content page for another one. Transaction identification assumes that user sessions have already been identified.

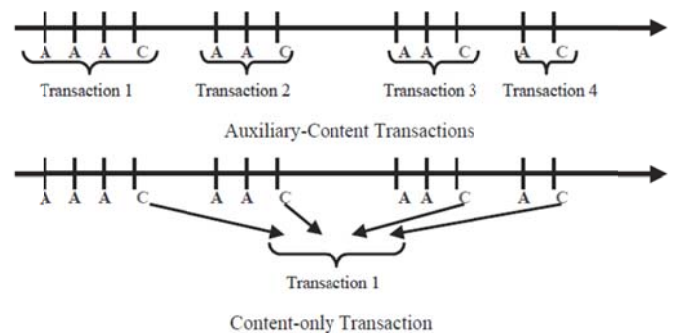


Figure 2. The session of individual user: the path through the navigation pages to the content page (Auxiliary-Content) and session focused merely on the content pages (Content-only) [15].

Using the concept of auxiliary and content page references, there are two ways to define transactions, as shown in Figure 2.

1. The first would be to define a transaction as all of the auxiliary references up to and including each content reference for a given user (Auxiliary – Content Transactions). Mining these auxiliary-content transactions would essentially give the common traversal paths through the web site to a given content page.

2. The second method (Content-only Transactions) would be to define a transaction as all of the content references for a given user. Mining these content-only transactions would give associations between the content pages of a site, without any information as to the path taken between the pages.

The first method of session is important for searching of user behavioral patterns as well as discovering errors in navigation or inaccuracies on the web pages.

We consider necessary to remark that the methods of navigation-driven user identification suppose that individual user pass through several navigation pages until she/he finally finds required content page. The found content page is the end of the session. The next searching of another content page using auxiliary pages is considered as a new session. The path through the auxiliary pages is often referred to as transaction in scientific literature.

V. THE METHODS OF NAVIGATION ORIENTED USER SESSION IDENTIFICATION

We describe two navigation oriented methods of user session identification. Both assume that two sets of transactions, namely auxiliary-content or content-only, can be formed. The first one, the maximal forward reference approach has an advantage over the reference length method in that it does not require an input parameter that is based on an assumption about the characteristics (type of statistical distribution) of a particular set of data [15].

A. Maximal Forward Reference Transaction Identification Method

The maximal forward reference transaction identification method is based on the work presented by Chen [20]. Instead of time spent on a web page, each transaction is defined as a set of web pages in the path from the first page in a user session up to the page before a backward reference is made. A forward reference is defined as a web page not already in the set of web pages for the current transaction.

Similarly, a backward reference is defined as a web page that is already contained in the set of web pages for the current transaction.

A new transaction is started when the next forward reference is made. The underlying model for this method is that the maximal forward reference pages are the content pages, and the web pages leading up to each maximal forward reference are the auxiliary pages. Like the reference length approach (a method described later), two sets of transactions, namely auxiliary-content or content-only, can be formed. The definition of a general transaction is used within the maximal forward reference method [15].

B. Reference Length Method

The reference length transaction identification method is based on the assumption that the amount of time a user spent on a web page correlates to whether the web page should be classified as an auxiliary or content page for that user. Qualitative analysis of several other server logs reveals that like Figure 3, the shape of the histogram has a large exponential component [15].

If we defined the assumption about the portion of navigation pages in surveyed log file, we can define the cut-off time C that separates the content pages and other types of pages.

When the cut-off time C is known the session can be created in such manner that we compare the time of particular web page visit with the cut-off time C . The session is then defined as a path through the navigation type of pages (duration of time spent on this web page is less than C) to the content page (the user spent there more time than C). We can claim the content page is the last page of session. The subsequent page is the first page of a new session.

VI. THE CALCULATION OF CUT-OFF TIME

The calculation of cut-off time C is the most important if we want to use the reference length transaction method for user session identification. The verification of exponential

distribution of variable $RLength$ obtained from the log file is also coessential. We took this assumption into consideration and verified it on the data obtained from the log file of university web site.

Analyzed records were created during the period of 12 days in October, 2011. Records were cleaned using conventional log file pre-processing methods. Redundant data about requesting other file types (.js, .css, .gif) as well as records about crawlers were removed. The final log file contained about 210 000 records that represent individual clicks of web site visitors.

In the next step, we analyzed the log file and we added the information about time spent by individual user on the web page into each record. This time was calculated as the difference of two consecutive records from one IP address. We did not realize the user session identification by the reason that we use the calculation of cut-off time C for reference length transaction identification.

Figure 3 depicts the histogram of the distribution of variable $RLength$ that represents time spent on the individual web pages.

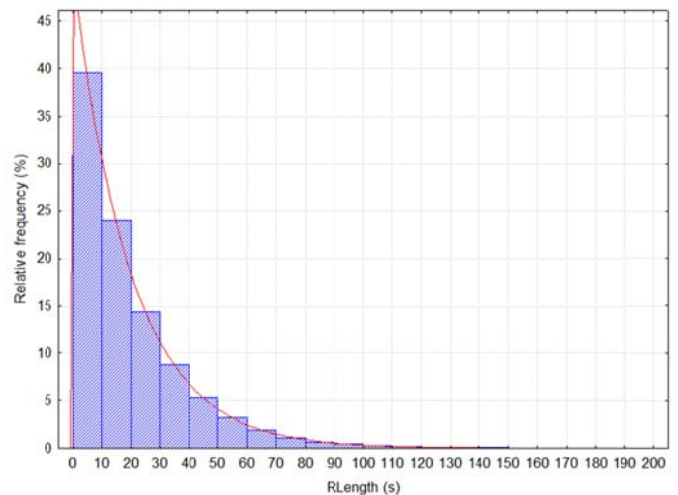


Figure 3. Exponential distribution of variable $RLength$.

We assume that the variance of the times spent on the auxiliary pages is small because the visitor only goes through them with the objective to find required information on content page. Therefore the auxiliary references make up the lower end of the curve (Figure 3). The variance of the times spent on the content pages is wide and we assume that they make up the upper tail that extends out to the longest reference.

If the assumption about the proportion of navigation pages in log file exists we can calculate the cut-off time C that divides web pages into navigation pages and content pages. We do not reject the null hypothesis which claims that the variable $RLength$ has assumed distribution (Figure 3).

The variable $RLength$ has exponential distribution.

$$f(RLength) = \lambda e^{-\lambda RLength}, \quad (9)$$

$$F(RLength) = 1 - e^{-\lambda RLength}, \quad (10)$$

where $RLength \geq 0$.

If p is relative frequency of navigation pages we can apply the fractile function (inverse distribution function) to estimate cut-off time C .

$$F^{-1}(p, \lambda) = C = \frac{-\ln(1-p)}{\lambda}, \quad (11)$$

where $0 \leq p < 1$.

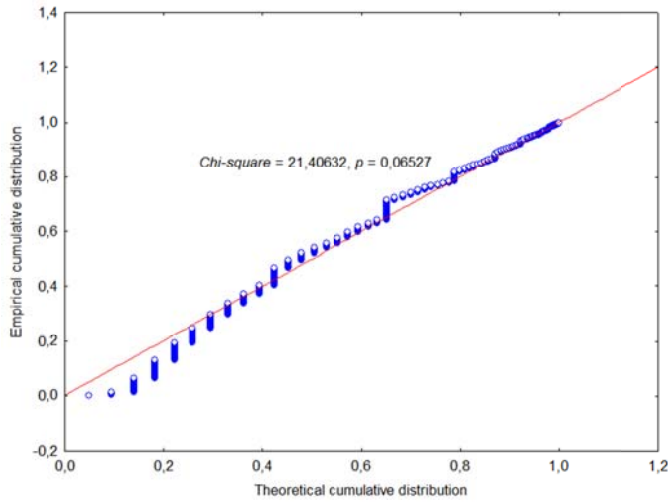


Figure 4. Probability of exponential distribution of variable $RLength$.

Maximum likelihood estimation of parameter λ (mean intensity of events) is

$$\hat{\lambda} = \frac{1}{\overline{RLength}}, \quad (12)$$

where $\overline{RLength}$ is observed mean length of visits. (Inverted value of mean time spent on the web pages).

Figure 5 describes cumulative distribution of time spent on the web pages expressed as a percentage. If the proportion of navigation pages were 70 % the cut-off time C would be between 20 and 30 seconds.

Once the cut-off time is estimated we can identify the session by comparison of each time with the cut-off time. The value of cut-off time divides the web pages into two groups – navigation and content pages.

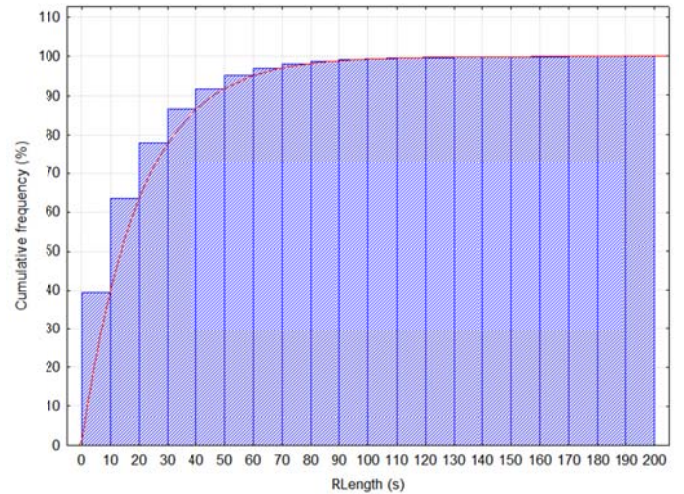


Figure 5. Cumulative frequencies of variable $RLength$.

If we have cut-off time estimated the session can be defined as a sequence of visited web pages with timestamp for which:

$$\langle \text{USID}, \langle \text{URL}_1, \text{DTime}_1, \text{RLength}_1 \rangle, \dots, \langle \text{URL}_k, \text{DTime}_k, \text{RLength}_k \rangle \rangle, \quad (13)$$

$$RLength_i \leq C, \quad (14)$$

where $1 \leq i < k$.

For the last web page of the session:

$$RLength_k > C, \quad (15)$$

The web page with the property (15) defines the next session. The first $k-1$ web pages are classified as navigation pages. The time spent on them is less or equal to cut-off time. The last k -th web page is classified as a content page. The time spent on this page is greater than the cut-off time.

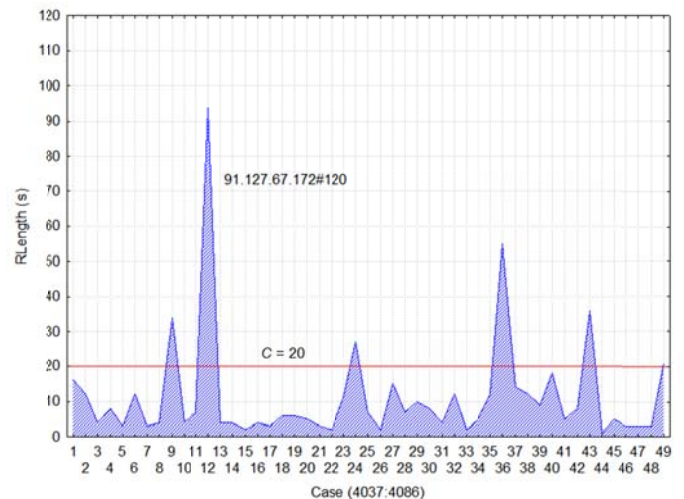


Figure 6. Session identification based on cut-off time.

The x-axis (Figure 6) depicts the sequence of web pages visited from particular IP address and agent. The sequence is ordered by the time. The y-axis represents the time spent on the web page. We estimated the value of variable C to 20 seconds.

The sequence of web pages with numbers 1 to 9 create the first session – pages 1 to 8 are navigation pages, the ninth web page is classified as a content web page. The next session is defined as the sequence of web pages with numbers 10 to 12 (Figure 6).

VII. DISCUSSION AND CONCLUSION

The length of each reference is estimated by taking the difference between the time of the next reference and the current reference. Obviously, the last reference in each transaction has no “next” time to use in estimating the reference length. The reference length method makes the assumption that all of the last references are content references, and ignores them while calculating the cut-off time. This assumption can introduce errors if a specific auxiliary page is commonly used as the exit point for a web site.

While interruptions such as a phone call or lunch break can result in the erroneous classification of an auxiliary reference as a content reference, it is unlikely that the error will occur on a regular basis for the same page. A reasonable minimum support threshold during the application of a data mining algorithm would be expected to weed out these errors.

We have to say that the assignment of particular page to the group of navigation or content pages may be different for each user. In order to group individual web page references into meaningful transactions for the discovery of patterns such as association rules, an underlying model of the user’s browsing behavior is needed.

For the purposes of association rule discovery, it is really the content page references that are of interest. The other page types are just to facilitate the browsing of a user while searching for required information, and will be referred to as auxiliary pages. What is merely an auxiliary page for one user may be a content page for another.

REFERENCES

- [1] J. Srivastava, R. Cooley, M. Deshpande and P. Tan, “Web usage mining: Discovery and applications of usage patterns from web data”, *ACM SIGKDD* Vol. 1 12-23, 2000.
- [2] D. Pierrakos, G. Paliouras, Ch. Papatheodorou, C. D. Spyropoulos, “Web usage mining as a tool for personalization: A survey”, *User Modeling and User-Adapted Interaction*, Kluwer Academic Publishers. Vol. 13, 311-372, 2003.
- [3] F. Tao, F. Murtagh, “Towards knowledge discovery from WWW log data” in *International Conference on Information Technology: Coding and Computing (ITCC'00)*, pp. 302-307, 2000.
- [4] Ch. Rana, “A study of web usage mining research tools”, *Int. J. Advanced Networking and Applications*, Vol. 3, pp. 1422-1429, 2012.
- [5] M. Koutri, N. Avouris, S. Daskalaki, “Chapter 7: A survey on web usage mining techniques for web-based adaptive hypermedia systems”, in S. Y. Chen and G. D. Magoulas (ed), *Adaptable and Adaptive Hypermedia Systems*, IRM Press, Hershey, pp. 125-149, 2005.
- [6] R. Kosala, H. Blockeel, “Web mining research: A survey”. *ACM SIGKDD*, 2000.

- [7] F. M. Facca, P. L. Lanzi, “Recent developments in web usage mining research”, *DaWak*, pp. 140-150, 2003.
- [8] F. M. Facca, P. L. Lanzi, “Mining interesting knowledge from weblogs: a survey”. *Data & Knowledge Engineering*, vol. 53, pp. 225-241, 2005.
- [9] Z. Chen, A. W. Fu, F. Ch. Tong, “Optimal algorithms for finding user access sessions from very large web logs” in *6th Pacific-Asia Conference on Advances in Knowledge Discovery and Data Mining*, Springer-Verlag, London, pp. 290-296, 2002.
- [10] C. Zhang, L. Zhuang, “New path filling method on data preprocessing in web mining” in *Proceedings of Computer and Information Science*, pp. 112-115, 2008.
- [11] Y. Li, B. Feng, O. Mao, “Research on path completion technique in web usage mining” in *International Symposium on Computer Science and Computational Technology*, Vol. 1, pp. 554-559, 2008.
- [12] V. Chitraa, A. S. Thanamani, “A novel technique for sessions identification in web usage mining preprocessing”, *International Journal of Computer Applications*, Vol. 34, no. 9, pp. 23-27, 2011.
- [13] G. Arumugam, “Optimal algorithms for generation of user session sequences using server side web user logs” in *International Conference on Network and Service Security*, pp. 1-6, 2009.
- [14] V. Chitraa, A. S. Davamani, “An efficient path completion technique for web log mining in IEEE International Conference on Computational Intelligence and Computing Research, 2010.
- [15] R. Cooley, B. Mobasher, J. Srivastava, “Knowledge and Information System”, 1. Springer-Verlag, 1999, ISSN 0219-1377.
- [16] B. Berendt, M. Spiliopoulou, “Analysis of navigation behaviour in web sites integrating multiple information systems.” in *VLDB Journal*, 9 (1), 2000, pp. 56-75.
- [17] M. Munk, M. Drlik, “Influence of different session timeouts thresholds on results of sequence rule analysis in educational data mining” in *Communications in Computer and Information Science*, 166 CCIS (PART 1), 2011, pp. 60-74.
- [18] M. Munk, J. Kapusta, P. Švec, M. Turčáni, “Data advance preparation factors affecting results of sequence rule analysis in web log mining” in *E & M Ekonomie a Management*, Vol. 13, Issue 4, 2010, 143-160, ISSN 1212-3609.
- [19] M. Munk, J. Kapusta, P. Švec, “Data preprocessing dependency for web usage mining based on sequence rule analysis” in *Proceedings of the IADIS European Conference on Data Mining 2009, ECDM'09, Part of the IADIS MCCSIS 2009*, 179-181, ISBN 978-972892488-1.
- [20] Chen, Ming-Syan, Park, Jong Soo, Yu, Philip S., “Data mining for path traversal patterns in a web environment.” In: *Proceedings - International Conference on Distributed Computing Systems*, 1996., pp. 385-392.
- [21] B. Mobasher, R. Cooley, J. Srivastava, “Automatic Personalization Based on Web Usage Mining”, *Communications of the ACM*, 43 (8), 2000, pp. 142-151.

Secure Data Aggregation in Wireless Multimedia Sensor Networks via Watermarking

Ersin Elbaşı
TUBITAK
Ankara, TURKEY
ersi.elbasi@tubitak.gov.tr

Suat Özdemir
Computer Engineering Department
Faculty of Engineering, Gazi University
Ankara, TURKEY
suatozdemir@gazi.edu.tr

Abstract— Wireless Multimedia Sensor Networks (WMSNs) consist of sensor nodes that have multiple sensing units. Unlike traditional wireless sensor networks, WMSNs are used to collect multimedia data such as video or image. As WMSNs are employed by military or civil surveillance/tracking applications security of these Networks must be ensured. As in traditional wireless sensor Networks, sensor nodes in WMSNs are resource limited. Hence, data collection in WMSNs must be performed in an energy efficient way. In this study, a watermarking based protocol that ensures secure and energy efficient data collection in WMSNs is proposed. In the proposed protocol, sensor nodes embed the secret data into image data using an attack resilient watermarking algorithm and data aggregators aggregate the image data. Performance analysis and simulations show that the proposed protocol ensures secure data aggregation and reduces the amount of data transmission.

Keywords- Heterogeneous wireless sensor networks, concealed data aggregation, security, watermarking

I. INTRODUCTION

Wireless sensor networks often consists of a large number of low-cost sensor nodes that have strictly limited sensing, computation, and communication capabilities. Due to resource restricted sensor nodes, it is important to minimize the amount of data transmission so that the average sensor lifetime and the overall bandwidth utilization are improved [1]. Wireless Multimedia Sensor Networks (WMSNs) are useful in event detection systems monitoring of the changes of physical phenomena in the surrounding environments [1, 2]. As military surveillance and health monitoring are the major application areas of WMSNs, securing these networks is another important issue [2].

Data aggregation is the process of summarizing and combining sensor data in order to not only to reduce the throughput of data transmission, thus saving energy effectively [3,4], but also to enhance the accuracy of event detection and avoid interference of the compromised nodes.

Although data aggregation and security are both indispensable tools for WMSNs, in terms of security, there is a significant risk of data aggregation. A sensor node that is compromised by an adversary can illegally disclose the data it collects from other nodes. If the compromised node is a data aggregator, the attack is more severe. For example, by capturing a few number of data aggregators that are positioned close to the base station, adversaries can attack on the security of the data of a large portion of the wireless sensor network. Hence, a security aware data aggregation protocol should keep sensor data secret from data aggregators, i.e., providing data privacy, while still allowing data aggregators to perform data aggregation.

In this paper, we propose a watermarking based protocol that ensures secure and energy efficient data collection in WMSNs. In the proposed protocol, sensor nodes embed the secret data into image data using an attack resilient watermarking algorithm and data aggregators aggregate the image data. Performance analysis and simulations show that the proposed protocol ensures secure data aggregation and reduces the amount of data transmission.

The rest of the paper is organized as follows. In Section 2, we present the state of the art in the area. In Section 3, system model is explained. The proposed protocol is explained in Section 4. Section 5 presents the simulation results. Finally, concluding remarks are given in Section 6.

II. RELATED WORK

In the wireless sensor network domain, the secure data aggregation has been extensively studied [4-15]. In [7], a security mechanism that detects node misbehaviors, such as dropping or forging messages and transmitting false data, is presented. In [8], random sampling mechanisms and interactive proofs are used to check the correctness of the aggregated data at base station. In [5], sensor nodes first send data aggregators the characteristics of their data to determine which sensor nodes have distinct data; those sensor nodes with distinct data then transmit their encrypted data. In [9], sensor nodes use the cryptographic algorithms only when cheating activity is detected. The authors of [10] propose that during a normal hop-by-hop aggregation process in a tree topology, more trust is placed on the high-level sensor nodes (i.e., nodes closer to the root) than the low-level nodes. [11] proposes a

protocol that makes use of a web of trust to overcome the shortcomings of cryptography based secure data aggregation solutions.

Digital watermarking has received increasing attention in recent years. Distribution of movies, music, and images is now faster and easier via computer technology, especially on the internet. There are two main methods in watermarking: spatial domain watermarking and transform domain watermarking. Least Significant Bit (LSB) embedding is the earliest and also the simplest technique. Since the last binary bits are the least significant bits, their modification will not be perceived by human eyes. Cox et al. [16] proposed secure spread spectrum watermarking algorithm. This algorithm uses the Discrete Cosine Transformation (DCT) in gray scale image. Wang et al [17] proposed a watermark embedding algorithm, which adds the watermark to the host in the spatial domain instead of the transform domain, reducing the complexity considerably. Piva et al. [18] presented DCT based watermark recovering without resorting to the uncorrupted original image. This algorithm provides extra robustness against intentional attacks and distortions. Dugad et al. [19] proposed wavelet based scheme for watermarking images. Caldelli et al. [20] proposed geometric invariant in Discrete Fourier Transformation (DFT) domain. Zhu et al. [21] proposed multi-resolution watermarking for image and videos.

III. SYSTEM MODEL

We consider a clustered static WMSN that consists of sensor nodes distributed over a predetermined area and a single base station. Sensor nodes have more than one sensing unit and each cluster has a data aggregator node. Data aggregators collect and compress the data of their surrounding sensor nodes. The compressed data are sent to the base station over a multi hop link. Sensor nodes have limited resources such as battery power and processors. Due to wireless medium transmitted data packets can be compromised by malicious users. Hence, data packets are encrypted and the encrypted packets can be decrypted by only the base station. Malicious users can also physically compromise the sensor nodes. This paper aims to prevent the compromised nodes to alter the data aggregation process.

A. Data aggregation

Simple data aggregation techniques such as max/min/average are not used in the aggregation of multimedia data. In this paper, secret data are embedded into images and therefore data aggregation takes place in the form of Huffman compression. In the proposed protocol, images are obtained in JPEG format, they are divided into blocks and discrete cosine transformation is applied to these blocks. Data aggregators compress the collected data blocks via Huffman coding.

B. Digital Watermarking

There are several criteria to classify watermarking techniques such as visible-invisible, blind-semi blind-non blind etc. Visible watermarks can be seen by eyes. For

example, the HBO logo on the screen is a visible watermark. If we record the show, the logo will be still on the screen. In invisible watermarking, the watermark is not visible at all. A watermarking technique that requires the original image to detect the watermark is called non-blind watermarking. A blind technique does not require the original image to detect watermark. Semi-blind watermarking techniques require the seed and the watermarked document for detection. Another criteria in watermarking is the watermark type: Visual watermark and PRN sequence. The PRN sequence allows the detector to statistically check the presence or absence of a watermark. A PRN sequence generated by feeding a linear or nonlinear generator with a secret key. However, embedding a meaningful watermark (visual watermark) is essential in most applications. This watermark might be a binary image, stamp, company logo or label [17-21]. The desired requirements from a watermarking technique are as follows:

1. The watermark should be secure. Detection or removal of the watermarks should be impossible.
2. The watermark detection should be reliable.
3. The watermarking algorithm should be resistant against all types of attacks.
4. The watermark should have a high data capacity.
5. The watermark should be transparent.

IV. THE PROPOSED PROTOCOL

This section explains the proposed protocol.

A. Data Collection

In the proposed protocol, sensor nodes collect both numeric data such as temperature, humidity and visual data such as images. We assume that only numeric data are secret. Images can be sent in the network without any encryption. Hence, each sensor node embeds its secret numeric data into its image data and sent this image to the data aggregator. The data embedding is explained next.

B. Data embedding and extracting via watermarking

To embed secret information to cover image first we convert RGB image to YUV image. The RGB color model is an additive model in which red, green and blue (often used in additive light models) are combined in various ways to reproduce other colors. The name of the model and the abbreviation "RGB" come from the three primary colors, Red, Green and Blue. The YUV model defines a color space in terms of one luminance and two chrominance components. Y stands for the luminance component (the brightness) and U and V are the chrominance (color) components.

$$\begin{aligned} R, G, B, Y &\in [0, 1] \\ U &\in [-0.436, 0.436] \\ V &\in [-0.615, 0.615] \\ Y &= 0.299R + 0.5876G + 0.111B \\ U &= -0.147R - 0.289G + 0.436B \\ V &= 0.615R - 0.515G - 0.100B \end{aligned}$$

Embedding and extraction algorithms are given below:

Embedding the secret information:

Input: Image I and binary visual watermark W.

Process:

- Convert the NxM RGB image to YUV image.
- Compute the DWT of the luminance layer Y.
- Modify the DWT coefficients V_{ij} in the LL, LH, HL, and HH bands.
- $V_{w,ij} = V_{ij}^k + \alpha_k \cdot W_{ij}$ where $i = 1, \dots, n; j = 1, \dots, m$
- Apply inverse DWT to obtain the watermark cover image I_w .

Output: Watermarked cover image

Extraction of the secret information:

Input: Watermarked (possibly attacked) image.

Process:

- Covert NxM RGB I image to YUV.
- Compute the DWT of the luminance layer Y.
- Extract the binary visual watermark from the LL,LH,HL and HH bands.
- $W_{ij}^* = (V_{w,ij}^{*k} - V_{ij}^k) / \alpha_k$ where $i = 1, \dots, n; j = 1, \dots, m$
- If $W_{ij}^* > T$, then $W_{ij}^* = 1$ else $W_{ij}^* = 0$, where T is the threshold between 0 and 1.

Output: Binary visual watermark (secret information).

C. Data Aggregation

For multimedia data aggregation, simple data aggregation techniques such as max/min/average cannot be not used. For image data, aggregation can be achieved via compression. The proposed protocol uses well-known data compression techniques known as Discrete Cosine Transformation (DCT) and Huffman coding [22]. Data aggregators divide each sensor node’s data into 4 blocks and apply DCT to each block. After DCT several sparse matrixes are obtained. Using K largest coding algorithm only the K-largest coefficients in each block are kept whereas the rest are discarded. Based on the K coefficients Huffman coding is applied to reduce the data transmission amount. The details of DCT and Huffman coding is explained in [22].

V. PERFORMANCE EVALUATION

In performance analysis, we evaluated the proposed protocol against the attacks on the images that have embedded secret information.

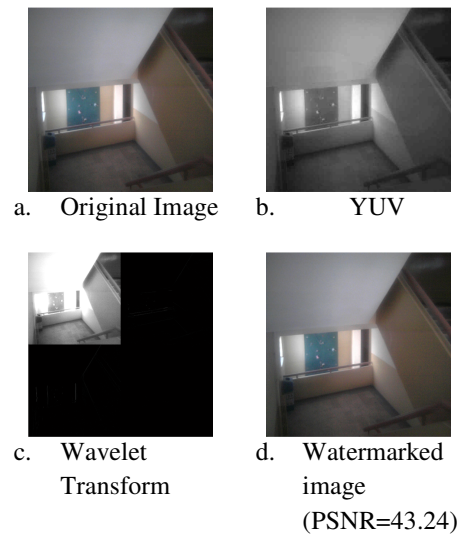


Figure 1. Original, Gray Scale, DWT and watermarked images

In figure 1 the original image, gray scale of the image, two level decomposition after wavelet transformation and watermarked image displayed. PSNR value of watermarked image shows that there is no difference between original image and watermarked image quality. The PSNR is most commonly used as a measure of quality of reconstruction in image watermarking. It is a ratio between the maximum value of a signal and the magnitude of background noise.

$$PSNR = 20 \times \log_{10} \left(\frac{255}{RMSE} \right)$$

where RMSE is the square root of MSE (mean square error). In figure 2 we show that watermarked data with PSNR values after some common attacks such as JPEG compression, JPEG 2000 compression, histogram equalization, low pass filtering, high pass filtering, resizing, Gauss noise, Gamma correction, rotation and contrast adjustment. This PSNR values shows that there is very small difference between watermarked images and attacked watermarked images.

In figure 3, extracted watermark (hidden data) and their similarity ratio after attacks has been given. Similarity Ratio (SR) Defined by $SR = S/(S+D)$, where S denotes the number of matching pixel values in compared images, and D denotes the number of different pixel values in compared images. The Similarity Ratio is used in evaluation of non-blind watermark extraction. Original image similarity ratio is 1 and similarity ratio after attacks (JPEG compression, JPEG 2000 compression, histogram equalization, low pass filtering, high pass filtering, resizing, Gauss noise, Gama correction, rotation and contrast adjustment) is between 0,54 and 0,73. These values shows that our data hiding algorithm is highly resist against most of the attacks.

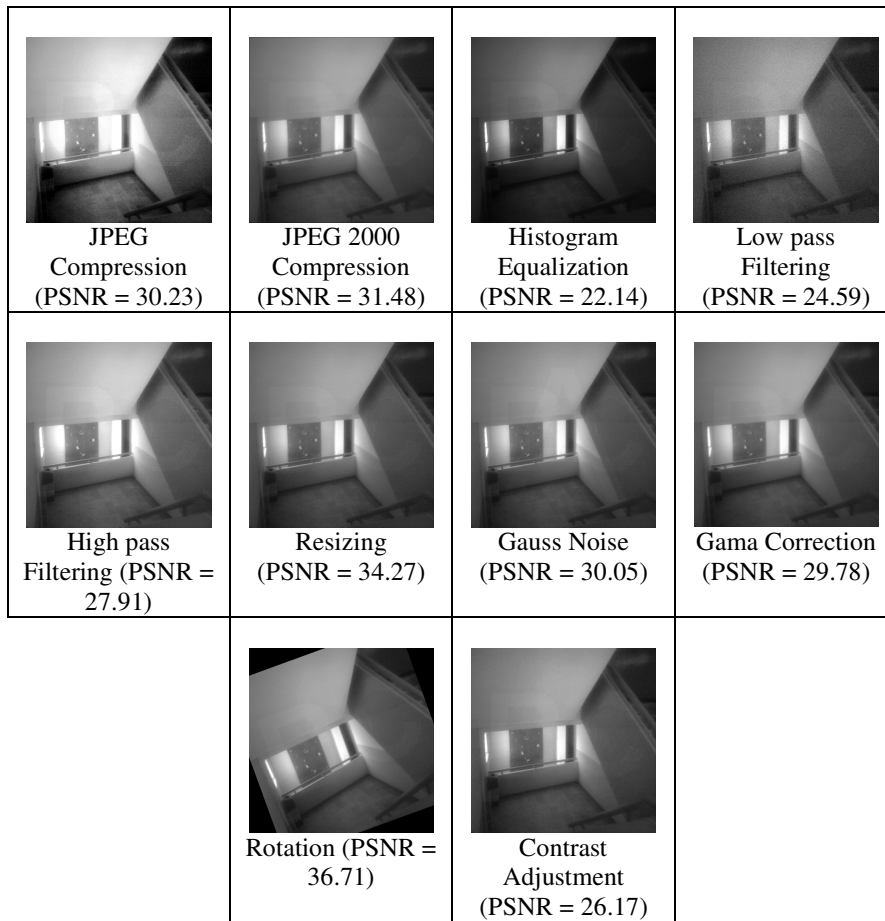


Figure 2. PSNR values after attacks in watermarked images

In figure 3, extracted watermark (hidden data) and their similarity ratio after attacks has been given. Similarity Ratio (SR) Defined by $SR = S/(S+D)$, where S denotes the number of matching pixel values in compared images, and D denotes the number of different pixel values in compared images. The Similarity Ratio is used in evaluation of non-blind watermark extraction. Original image similarity ratio is 1 and similarity ratio after attacks (JPEG compression, JPEG 2000 compression, histogram equalization, low pass filtering, high pass filtering, resizing, Gauss noise, Gama correction, rotation and contrast adjustment) is between 0,54 and 0,73. These values shows that our data hiding algorithm is highly resist against most of the attacks.

In figure 4 PSNR values given after common attacks with embedding %100, %75, %50 and %25 of the data. In figure 5 Similarity Ratios has been given after attacks. In figure 6 PSNR values has been given for scaling factor between 0 and 1. Highest scaling PSNR value in scaling factor 0,3 has been used in our embedding algorithm.

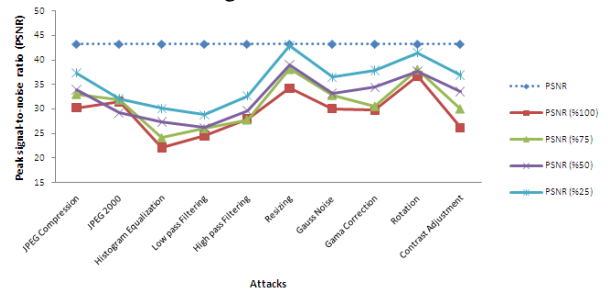


Figure 4. Image quality metrics of embedded different size of digital secret data

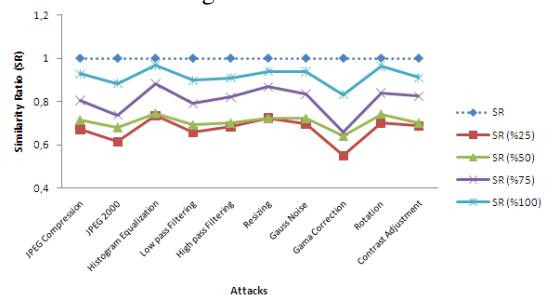


Figure 5. SR values of embedded different size of digital secret data

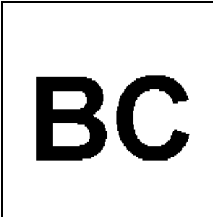
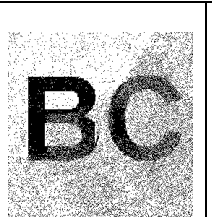
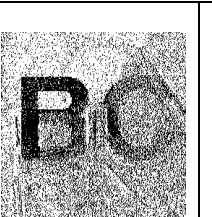
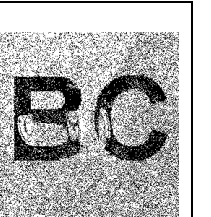
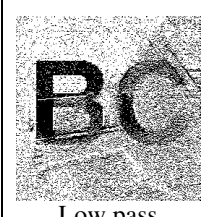
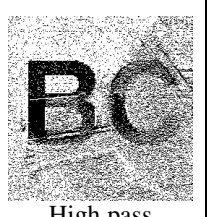
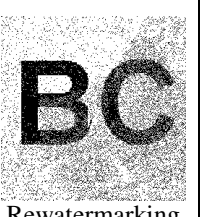
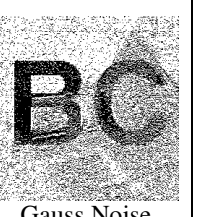
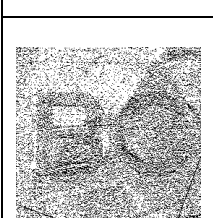
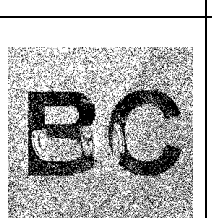
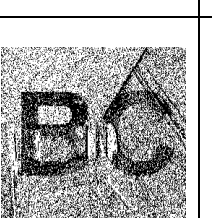
 <p>Original (SR=1.0000)</p>	 <p>JPEG Compression (SR=0.6712)</p>	 <p>JPEG 2000 Compression (SR=0.6147)</p>	 <p>Histogram Equalization (SR=0.7352)</p>
 <p>Low pass Filtering (SR=0.6594)</p>	 <p>High pass Filtering (SR=0.6841)</p>	 <p>Rewatermarking (SR=0.7231)</p>	 <p>Gauss Noise (SR=0.6956)</p>
 <p>Gama Correction (SR=0.5497)</p>	 <p>Rotation (SR=0.7002)</p>	 <p>Contrast Adjustment (SR=0.6874)</p>	

Figure 3. Extracted watermark (logo) and similarity ratios after attacks

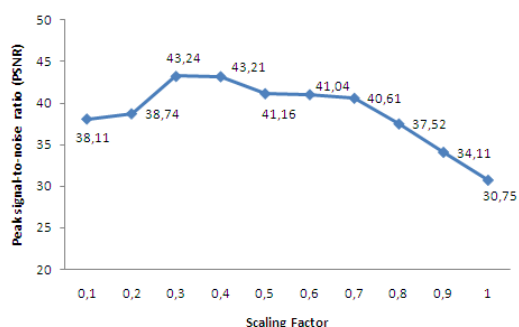


Figure 6. PSNR values after watermarking in different scale coefficients

VI. CONCLUSION

In this paper, we propose a secure data aggregation protocol for WMSNs based on digital watermarking techniques. In the proposed protocol, sensor nodes embed the secret data into image data using an attack resilient watermarking algorithm

and data aggregators aggregate the image data. Performance analysis and simulations show that the proposed protocol ensures secure data aggregation and reduces the amount of data transmission.

VII. REFERENCES

- [1] Akyildiz, I.F., Su, W., Sankarasubramaniam, Y. ve Cayirci, E., "A survey on sensor networks", *IEEE Communications Magazine*, 40(8), 102-114, 2002.
- [2] Ozdemir, S. ve Xiao, Y., "Secure Data Aggregation in Wireless Sensor Networks: A Comprehensive Overview", *Computer Networks*, 53(12), pp. 2022-2037, 2009.
- [3] Akyildiz, I.F., Melodia, T. ve Chowdhury, K. R., "A survey on wireless multimedia sensor Networks", *Computer Networks*. 51, 4, 921-960, 2007.
- [4] Cam, H., Ozdemir, S., Nair, P., Muthuavinashiappan, D. ve Sanli, H.O., "Energy-Efficient and secure pattern based data aggregation for wireless sensor Networks", *Computer Communications*, 29(4), 446-455, 2006.
- [5] Cam, H., Ozdemir, S., Nair, P. ve Muthuavinashiappan D., "ESPDA: Energy-Efficient and Secure Pattern-Based Data Aggregation for Wireless Sensor Networks", *Proc.*

- of IEEE Sensors 2003 Conference, Toronto, Canada, 2003.
- [6] Lee, S. ve Chung, T., "Data Aggregation for Wireless Sensor Networks Using Self organizing Map", *Artificial Intelligence and Simulation*, V. 3397, 508-517, 2005.
- [7] Hu, L. ve Evans, D., "Secure aggregation for wireless Networks", *Workshop on Security and Assurance in Ad hoc Networks*, 384-392, 2003.
- [8] Przydatek, B., Song, D. ve Perrig, A., "SIA : Secure information aggregation in sensor Networks", *SenSys'03*, 255 – 265, 2003.
- [9] Girao, J., Westhoff, D. ve Schneider, M., "Concealed Data Aggregation for Reverse Multicast Traffic in Sensor Networks: Encryption, Key Distribution, and Routing Adaptation", *IEEE Transactions on Mobile Computing*, 1417-1431, 2006.
- [10] Domingo-Ferrer, J., "A provably secure additive and multiplicative privacy homomorphism", *Information Security Conference, LNCS 2433*, 471-483, 2002.
- [11] Ozdemir, S., "Concealed Data Aggregation in Heterogeneous Sensor Networks using Privacy Homomorphism", *Proc. of ICPS 2007 : IEEE International Conference on Pervasive Services*, Istanbul, Turkey.
- [12] Castelluccia, C., Mykletun, E. ve Tsudik, G., "Efficient aggregation of encrypted data in wireless sensor Networks", *Conference on Mobile and Ubiquitous Systems: Networking and Services*, vol., no., pp. 109-117, 2005.
- [13] Ozdemir, S., "Secure Data Aggregation in Wireless Sensor Networks via Homomorphic Encryption (manuscript in Turkish)", *Journal of The Faculty of Engineering and Architecture of Gazi University*, vol.23, no.2, pp. 365-373, 2008.
- [14] Ozdemir S. ve Xiao. Y.. "Integrity Protecting Hierarchical Concealed Data Aggregation for Wireless Sensor Networks", *Computer Networks*, vol. 55, no. 8, pp. 1735-1746, 2011.
- [15] Ozdemir S. ve Xiao. Y.. "Polynomial Regression Based Secure Data Aggregation for Wireless Sensor Networks" *IEEE GLOBECOM 2011*, Dec. 5-9, Houston, TX.
- [16] Cox I. J., Kilian J., Leighton T. ve Shamoon T.. "Secure Spread Spectrum Watermarking for Multimedia." *IEEE Transactions on Image Processing*, 6(12), 1997, pp. 1673-1687.
- [17] Piva A., Barni M., Bartolini F. ve Cannellini F.. "DCT-based Watermark Recovering without Resorting to the Uncorrupted Original Image. Proceedings of the 1997 International Conference on Image Processing (ICIP '97), Washington, DC, USA., 1997.
- [18] Dugad R., Ratakonda K. ve Ahuja N.. "A New Wavelet-Based Scheme for Watermarking Images." *International Conference on Image Processing (ICIP 1998)*, Vol. 2, Chicago, IL, 1998, pp. 419-423.
- [19] Zhu W., Xiong Z. ve Zhang Y.Z.. "Multiresolution Watermarking for Images and Video." *IEEE Transactions on Circuits and Systems for Video Technology*, 9(4), 1999, pp. 545-550.
- [20] Elbasi E. ve Eskicioglu A. M.. "A DWT-based Robust Semi-blind Image Watermarking Algorithm Using Two Bands." *IS&T/SPIE's 18th Symposium on Electronic Imaging, Security, Steganography, and Watermarking of Multimedia Contents VIII Conference*, San Jose, CA, 2006.
- [21] Tao P. ve Eskicioglu A.M. "A Robust Multiple Watermarking Scheme in the DWT Domain." *Optics East 2004 Symposium. Internet Multimedia Management Systems V Conference*, Philadelphia, PA, 2004, pp. 133-144.
- [22] Huffman D.A.. "A Method for the Construction of Minimum-Redundancy Codes", *Proceedings of the I.R.E.*, , pp 1098-1102, 1952.
- [23] [Tiny OS Simulator. <http://www.tinyos.net/>.

Embedded Solution for Road Condition Monitoring Using Vehicular Sensor Networks

Artis Mednis*[†], Atis Elsts*[†], Leo Selavo*[†]

*Cyber-Physical Systems Laboratory
Institute of Electronics and Computer Science
14 Dzerbenes Str., Riga, LV 1006, Latvia

[†]Faculty of Computing
University of Latvia
19 Raina Blvd., Riga, LV 1586, Latvia
Email: {firstname.lastname}@edi.lv

Abstract—The advantages of vehicular sensor networks over common wireless sensor networks include the possibility to cover wide measurement area using relatively small number of sensor nodes as well as not so strong limitations according device dimensions, weight and power consumption. These attractive features are reason for expansion of vehicular sensor networks for various environmental monitoring tasks - from defects of road surface to air quality in urban areas.

The contribution of this paper is a customized embedded device dedicated for monitoring of road surface using microphone and accelerometer sensors as well as collection of meteorological data for creation of detailed road meteorology maps. Selected hardware and software aspects are discussed and the implementation of previously developed method for road surface monitoring using accelerometer data is presented.

Index Terms—road surface analysis, potholes, embedded device, vehicular sensor network

I. INTRODUCTION

A sensor is a device used for the measurement of some physical quantity or physical state. Common result of this measurement is an analog signal that is converted to a digital signal using ADC and subsequently processed using some computing device. Typically there is a need to make measurements in several locations and therefore a number of sensors should be configured, deployed and serviced if necessary. This approach - usage of many static deployed sensors - has its drawbacks as sooner or later scalability and maintenance issues will arise. To overcome this problem a large number of static deployed sensors could be substituted by few mobile sensors. One type of objects that could be used as mobile sensor carriers are vehicles. Among the possibility to cover wide measurement area there are other advantages including energy utilization from sensor carrier as well as not so strong limitations according device dimensions, weight and power consumption.

There are several categories of data that could be acquired using vehicles as sensor carriers. First of them is data about vehicle itself, for example, driving speed and actual location. Next data source is vehicle driver characterized by its pulse and time of the reaction. Environment measurements could

be collected, for example, data about acoustic noise and air pollution. Last but not least - vehicles are driving using specially deployed infrastructure including road surface, and regular monitoring of this infrastructure could help to optimize necessary maintenance works.

The main purpose of the CarMote embedded device described in this paper is monitoring of road surface using microphone and accelerometer sensors as well as collection of meteorological data for creation of detailed road meteorology maps. This research was inspired by successful verification of previously developed methods for road surface monitoring using general purpose computing devices and Android smartphones as well as by challenging task - implementation of these methods using customized embedded device.

Related work is discussed in Section II. Requirements and assumptions are listed in Section III. Proposed customized embedded device hardware and software is described and analyzed in Section IV. The evaluation of the developed embedded device including successful test in acquiring of road surface data from accelerometer sensor is described in Section V. The final section presents the conclusion that the proposed customized embedded device allows the implementation of previously developed method for road surface monitoring using accelerometer data and therefore demonstrates the suitability of the device in the context of vehicular sensor networks.

II. RELATED WORK

By our knowledge the term "vehicular sensor networks" is introduced in 2006 when researchers from University of California and University of Bologna declared a new network paradigm - the use of vehicles as sensors [1]. This paradigm was characterized by high computation power and high storage space therefore potential costs for network deployment and maintenance could be relative high. As primary application of vehicular sensor network was declared urban monitoring, for example, imaging of streets, recognizing of license plates and diffusing of relevant notifications to drivers or police agents [2]. Other applications developed by other researchers

include monitoring of infrastructure items as road surface [3], collection of real time vehicular parking information [4], measuring air quality in city areas [5] and mobile surveillance [6].

Among communication between vehicles (V2V or vehicle-to-vehicle) vehicular sensor networks could include communication between vehicles and Road Side Units (V2I or vehicle-to-infrastructure) [7] [8]. In this case, as the number of network nodes could be very large, an effective identity verification should be ensured [9]. In the contrast of traditional wireless sensor networks where network nodes are placed in static locations vehicular sensor networks are characterized by dynamic changes in network topology. Therefore best possible connectivity could be achieved using appropriate combinations of transmission time and transmission range [10]. Nevertheless data gathering using these dynamic networks and data muling and multi-hop forwarding strategies is supposed to have specific delays [11].

Our proposed CarMote embedded device assumes usage of relatively low computation power and low storage space that is characteristic for common wireless sensor network nodes. Combination of these aspects with vehicles as sensor carriers allows performing tasks where a large number of low cost deployed and maintained network nodes have advantages over a small number of high cost deployed and maintained network nodes.

III. REQUIREMENTS

The following requirements were chosen as a basis for the development of the first prototype of the CarMote embedded device:

- 1) The hardware of the first prototype should be based on MCU that is inexpensive and relatively widespread used including applications in the domain of wireless sensor networks. Selection of an advanced MCU best suited for each possible application scenario is left for future work.
- 2) The sensing part of the first prototype should include microphone that allows implementation of the RoadMic approach [12], accelerometer that allows implementation of the Potroid approach [13] and a set of meteorological sensors that allows making of experiments in creation of detailed road meteorology maps. There should be a possibility to add position metadata using GPS and a possibility to extend sensing part with other sensors in the future.
- 3) The power supply part of the first prototype should ensure the possibility to use electrical system of the vehicle as main power source and internal battery pack as alternatively power source with automatic switching between them. Implementation of internal battery charging as well as energy harvesting is left for future work.
- 4) The storage part of the first prototype should ensure the possibility to store acquired sensor data and corresponding position metadata on media that has a relatively widespread use and could be removed from the

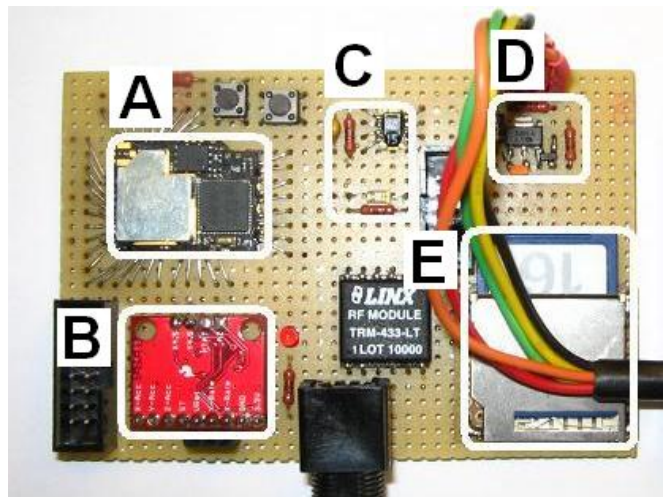


Fig. 1. The main board of the CarMote embedded device: A - TMote Mini wireless sensor network module, B - IMU Analog Combo Board, C - sensors SHT15 and TEMA6000, D - voltage regulators, E - SD flash memory card.

embedded device for data reading. Implementation of several supported media types as well as management of several media devices is left for future work.

- 5) The communication part of the first prototype should ensure the possibility to transmit acquired sensor data and corresponding position metadata using relatively widespread used communication standards. Usage of the communication part for other scenarios as investigation of the communication infrastructure is left for future work.
- 6) The software part of the first prototype should be based on operating system intended for usage in wireless sensor networks. The possibility to program several application scenarios without deep programming experience should be classified as advantage.

IV. APPROACH

The development of the first prototype of the CarMote embedded device was performed using as the basis the first version of LynxNet collar device developed during our past research activities related to wild animal monitoring using sensor networks [14]. This approach already fulfilled a part of previous set requirements. In addition, common hardware basis for both device types facilitates reusability of the software.

A. CarMote hardware design

The heart of the first prototype of the CarMote embedded device is TMote Mini wireless sensor network module (Fig. 1 - A). This module contains TI MCU MSP430F1611, TI/Chipcon transceiver CC2420 and ST EEPROM M25P80. The same MCU is used in other wireless sensor network modules, for example, EPIC [15], 3MATE! [16] and others. Selection of this popular MCU device allows to use the experience from previous developments as well as compatibility with available open source software.

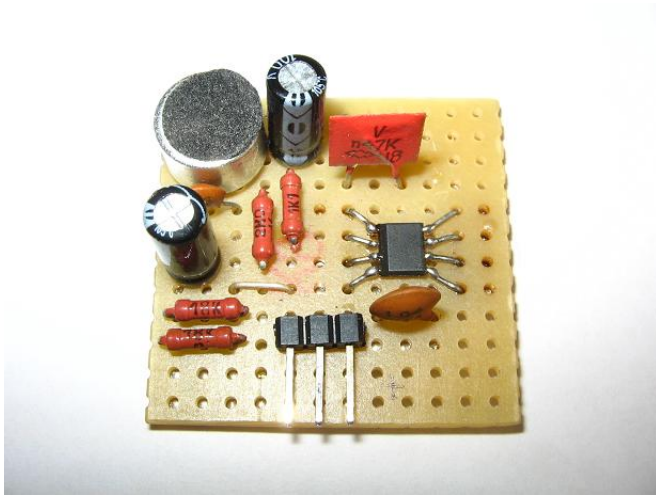


Fig. 2. The microphone board of the CarMote embedded device

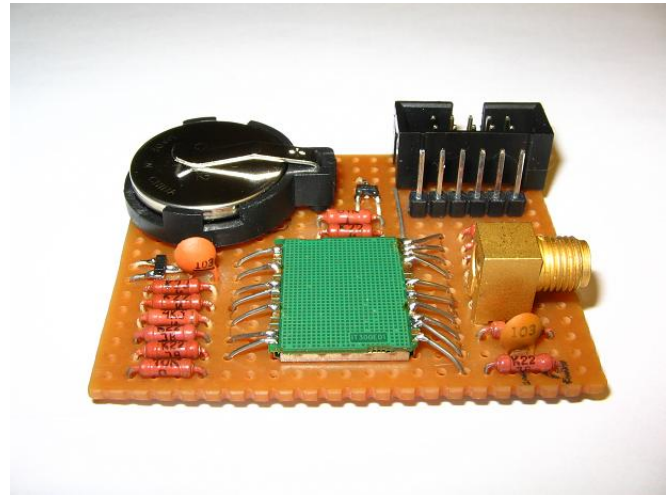


Fig. 3. The GPS board of the CarMote embedded device

To ensure the possibility of implementation of the Road-Mic approach the main board was equipped with attachable board consisting of electret microphone BCM9765P-44 and corresponding amplifier stage built using operational amplifier TS952ID (Fig. 2). The output of the amplifier stage is connected to the ADC input #7 of the MCU.

To ensure the possibility of implementation of the Potroid approach the main board was equipped with IMU Analog Combo Board from SparkFun (Fig. 1 - B). This board contains triple axis accelerometer ADXL335 and dual-axis gyroscope IDG500. Outputs of accelerometer X, Y and Z axis are connected to the ADC inputs #0, #1 and #2, but outputs of gyroscope axis X and Y - to the ADC inputs #3 and #4 of the MCU.

To ensure the possibility of experiments in creation of detailed road meteorology maps the main board was equipped with humidity and temperature sensor SHT15 as well as light sensor TEMT6000 (Fig. 1 - C). First of them was connected to MCU using I2C interface, but second - using ADC input #5. Implementation of barometric pressure sensor SCP1000 is left for future work.

To ensure the possibility to add position metadata using GPS the main board was equipped with attachable board consisting of GPS module Fastrax IT300 (Fig. 3). This board is connected to MCU using USART #0 interface and NMEA protocol. Operation of this board using SiRF protocol supported by Fastrax module is left for future work.

To ensure the possibility to use electrical system of the vehicle as main power source two sequential voltage regulators were used (Fig. 1 - D). First of them is dedicated to acquire 5V but second one - 3.3V. Four AA size battery pack is used as internal power source. Automatic switching between external power source and internal power source is ensured using low-loss Schottky diodes.

To ensure the possibility to store acquired sensor data and corresponding position metadata on media SD flash memory card was used (Fig. 1 - E). This removable media is connected

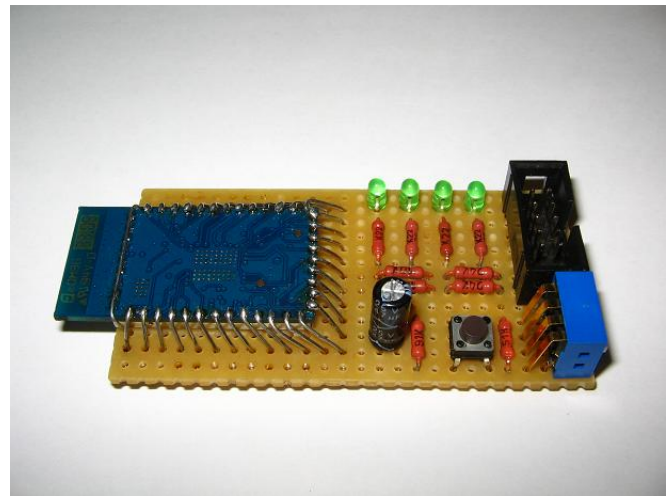


Fig. 4. The Wi-Fi board of the CarMote embedded device

to the MCU using SPI mode and respectively configured USART #1 interface.

To ensure the possibility to transmit acquired sensor data and corresponding position metadata two options were selected. First of them is Wi-Fi that could be used for medium range communication and the second one - Bluetooth that could be used for short range communication. Hardware for Wi-Fi communication was implemented as attachable board consisting of Roving Networks module RN-131C (Fig. 4) but hardware for Bluetooth communication - as attachable board consisting of Rayson module BT-220A2 (Fig. 5). Both attachable boards have serial interface for communication with MCU. During device prototyping stage just one module with serial interface (GPS, Wi-Fi, Bluetooth) is connected to MCU interface USART #0 at the time. Software driven multiplexer for commutation of several modules is left for future work.

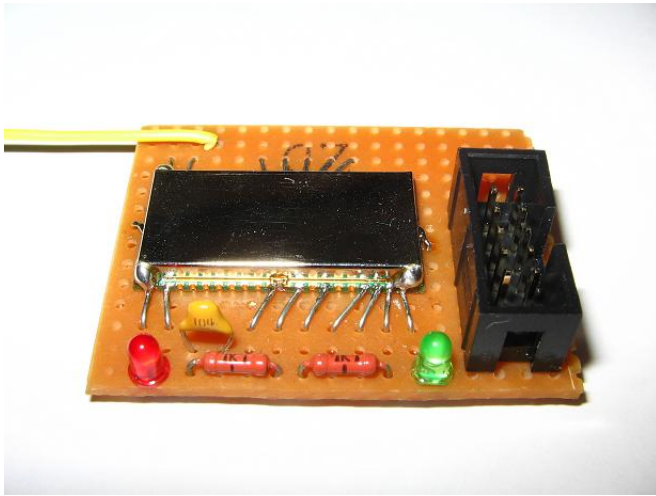


Fig. 5. The Bluetooth board of the CarMote embedded device

B. CarMote software design

In order to program CarMote, we have adapted MansOS operating system [17], a small and energy constrained device OS developed at the University of Latvia and Institute of Electronics and Computer Science (EDI). The OS aims to be user-friendly and easy to learn for individuals with C and UNIX programming experience.

The MCU of CarMote has a built-in 12-bit ADC. We sampled accelerometer's Z channel to evaluate the sampling speed. Sample rate 1820 samples per second (sps) was achieved without logging the data, and 1400 sps when logging the data to SD card. The rates are sufficient for the kind of applications CarMote was designed for, such as pothole detection.

An essential component of a highly mobile device is localization system. MansOS supports data interface with GPS devices, including NMEA¹ protocol parsing and online processing.

Last but not least, SD card support is included. It can be used either in *raw* mode, or by writing data to a filesystem. We have developed a custom filesystem to efficiently use local storage devices; although it is primary targeted for flash chips with no automatic rewrite options, it can be used for SD card as well. The FS provides buffering and automatic error detection features.

MansOS configuration system can enable these components when needed, or disable them to optimize compilation length. In either case, the system automatically detects when a component is not used and optimizes binary code by pruning unused components from the final executable.

A declarative scripting language called SEAL is available on top of MansOS. SEAL is targeted towards domain experts and novice programmers. SEAL features extremely concise syntax for describing common tasks: sensor sampling, data processing, and data communication. A few complete application examples are given in Listing 1 and Listing 2.

¹www.gpsinformation.org/dale/nmea.htm

Listing 1 SEAL code for accelerometer sampling and measurement storing

```
// define sensors
define AccelX AnalogIn, channel 0;
define AccelY AnalogIn, channel 1;
define AccelZ AnalogIn, channel 2;
// sample the sensors
read AccelX; read AccelY; read AccelZ;
// store sampled data to SD card
output LocalStorage;
```

Listing 2 SEAL code for pothole detection with STDEV algorithm

```
// define sensors
const ACCEL_Z 2;
define AccelZ AnalogIn, channel ACCEL_Z;
define Deviation stdev(take(AccelZ, 10));
// when STDEV value exceeds threshold:
when Deviation > 100:
    // read the detection time
    read Uptime;
    // indicate a pothole presence via beep
    use Beeper, on, duration 200ms, frequency 1000;
end
// log the detection time to SD card
output LocalStorage (Uptime);
```

V. EVALUATION

To evaluate the described CarMote embedded device the following set of the activities were performed:

- 1) test drive with Android smartphone HTC Desire and CarMote embedded device;
- 2) acquisition of the accelerometer data for pothole detection using Potroid approach;
- 3) comparative analysis of acquired accelerometer data.

Accelerometer data acquisition was performed 37 times per second using CarMote embedded device (Fig. 6) and 53 times per second using Android smartphone (Fig. 7). Analysis of the acquired data revealed that, taking into account slightly different positioning of both data collection devices, acquired data are practically identical and therefore data from CarMote embedded device are suitable for usage for pothole detection using Potroid approach. Relatively better sensitivity of the CarMote embedded device in the context of accelerometer Z axis value could be considered as advantage because this axis value is the most affected by potholes passed by vehicle.

Serious advantage of CarMote embedded device over Android smartphone equipped according Potroid approach and laptop computer equipped according RoadMic approach is the native possibility to use electrical system of the vehicle as main power source. In this case long term data acquisition and processing sessions are possible almost eliminating the hazard of empty internal power source.

VI. CONCLUSION AND FUTURE WORK

This paper describes CarMote embedded device dedicated for monitoring of road surface using microphone and accelerometer sensors as well as collection of meteorological

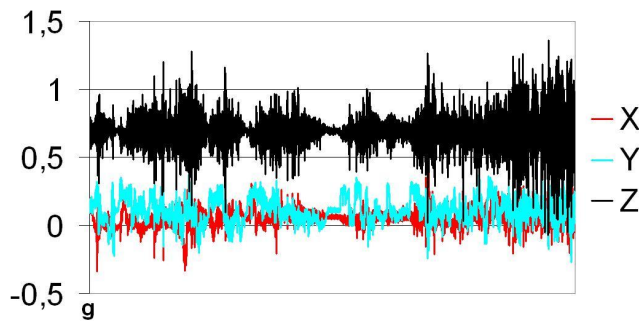


Fig. 6. Accelerometer data acquired using CarMote embedded device (a fragment, sampling rate 37 Hz)

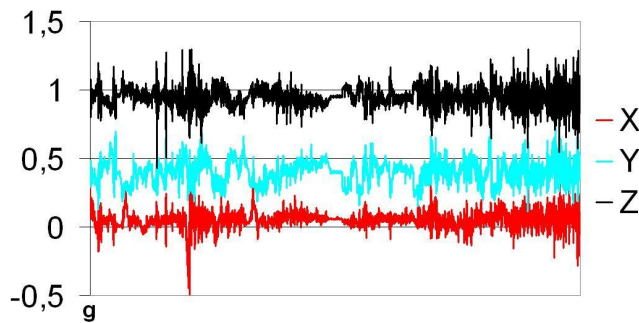


Fig. 7. Accelerometer data acquired using Android smartphone HTC Desire (a fragment, sampling rate 53 Hz)

data for creation of detailed road meteorology maps and its evaluation on a particular task - implementation of previously developed method for road surface monitoring using accelerometer data. The evaluation tests resulted in identical software operation and data acquisition in both hardware platforms - Android smartphone as well as CarMote embedded device. Therefore CarMote embedded device is suited for usage in vehicular sensor networks.

The future work includes development of the software for transmission of acquired sensor data and corresponding position metadata using Wi-Fi and Bluetooth as well as experiments in creation of detailed road meteorology maps.

VII. ACKNOWLEDGMENT

This work has been supported by the European Social Fund within the project "Support for Doctoral Studies at University of Latvia - 2" and Latvian National Research Program "Development of innovative multi-functional material, signal processing and information technologies for competitive and research intensive products". Special thanks to our IECS colleagues Girts Strazdins, Reinholds Zviedris, Georgijs Kanonirs and Andris Gordjusins for their help during hardware and software development as well as real-world experiments.

REFERENCES

[1] U. Lee, E. Magistretti, B. Zhou, M. Gerla, P. Bellavista, and A. Corradi, "Efficient data harvesting in mobile sensor platforms," in *Proceedings of the 4th annual IEEE international conference on Pervasive Computing*

and *Communications Workshops*, ser. PERCOMW '06. Washington, DC, USA: IEEE Computer Society, 2006, pp. 352–. [Online]. Available: <http://dx.doi.org/10.1109/PERCOMW.2006.47>

[2] U. Lee, B. Zhou, M. Gerla, E. Magistretti, P. Bellavista, and A. Corradi, "Mobeyes: smart mobs for urban monitoring with a vehicular sensor network," *Wireless Commun.*, vol. 13, no. 5, pp. 52–57, Oct. 2006. [Online]. Available: <http://dx.doi.org/10.1109/WC-M.2006.250358>

[3] J. Eriksson, L. Girod, B. Hull, R. Newton, S. Madden, and H. Balakrishnan, "The pothole patrol: using a mobile sensor network for road surface monitoring," in *Proceedings of the 6th international conference on Mobile systems, applications, and services*, ser. MobiSys '08. New York, NY, USA: ACM, 2008, pp. 29–39. [Online]. Available: <http://doi.acm.org/10.1145/1378600.1378605>

[4] S. Mathur, S. Kaul, M. Gruteser, and W. Trappe, "Parknet: a mobile sensor network for harvesting real time vehicular parking information," in *Proceedings of the 2009 MobiHoc S3 workshop on MobiHoc S3*, ser. MobiHoc S3 '09. New York, NY, USA: ACM, 2009, pp. 25–28. [Online]. Available: <http://doi.acm.org/10.1145/1540358.1540367>

[5] S.-C. Hu, Y.-C. Wang, C.-Y. Huang, and Y.-C. Tseng, "Measuring air quality in city areas by vehicular wireless sensor networks," *J. Syst. Softw.*, vol. 84, no. 11, pp. 2005–2012, Nov. 2011. [Online]. Available: <http://dx.doi.org/10.1016/j.jss.2011.06.043>

[6] C.-M. C. Kun-chan Lan and H.-Y. Wang, "Using vehicular sensor networks for mobile surveillance," September 2012.

[7] M. J. Piran, G. R. Murthy, and G. P. Babu, "Vehicular ad hoc and sensor networks: principles and challenges," *CoRR*, vol. abs/1108.2776, 2011. [Online]. Available: <http://dblp.uni-trier.de/db/journals/corr/corr1108.html#abs-1108-2776>

[8] P. R. K. Reddy, P. Joshna, and G. Sireesha, "Data collection through vehicular sensor networks," *CoRR*, vol. abs/1206.6281, 2012.

[9] C. Zhang, R. Lu, X. Lin, P.-H. Ho, and X. Shen, "An efficient identity-based batch verification scheme for vehicular sensor networks," in *INFOCOM*. IEEE, 2008, pp. 246–250. [Online]. Available: <http://dblp.uni-trier.de/db/conf/infocom/infocom2008.html#ZhangLLHS08>

[10] H. Conceição, M. Ferreira, and J. a. Barros, "On the urban connectivity of vehicular sensor networks," in *Proceedings of the 4th IEEE international conference on Distributed Computing in Sensor Systems*, ser. DCOSS '08. Berlin, Heidelberg: Springer-Verlag, 2008, pp. 112–125. [Online]. Available: http://dx.doi.org/10.1007/978-3-540-69170-9_8

[11] C. E. Palazzi, F. Pezzoni, and P. M. Ruiz, "Delay-bounded data gathering in urban vehicular sensor networks," *Pervasive Mob. Comput.*, vol. 8, no. 2, pp. 180–193, Apr. 2012. [Online]. Available: <http://dx.doi.org/10.1016/j.pmcj.2011.06.008>

[12] A. Mednis, G. Strazdins, M. Liepins, A. Gordjusins, and L. Selavo, "Roadmic: Road surface monitoring using vehicular sensor networks with microphones," in *NDT (2)*, ser. Communications in Computer and Information Science, F. Zavoral, J. Yaghob, P. Pichappan, and E. El-Qawasmeh, Eds., vol. 88. Springer, 2010, pp. 417–429. [Online]. Available: <http://www.springerlink.com/content/q3t5564544t8x188>

[13] A. Mednis, G. Strazdins, R. Zviedris, G. Kanonirs, and L. Selavo, "Real time pothole detection using android smartphones with accelerometers," in *DCOSS*. IEEE, 2011, pp. 1–6.

[14] R. Zviedris, A. Elsts, G. Strazdins, A. Mednis, and L. Selavo, "Lynxnet: Wild animal monitoring using sensor networks," in *REALWSN*, ser. Lecture Notes in Computer Science, P. J. Marrón, T. Voigt, P. I. Corke, and L. Mottola, Eds., vol. 6511. Springer, 2010, pp. 170–173.

[15] Prabal Dutta, *Epic: An Open Mote Platform for Application-Driven Design*, March 2007.

[16] TRETEC S.r.l., *3Mate! - Wireless Sensor Module*, July 2011.

[17] G. Strazdins, A. Elsts, and L. Selavo, "Mansos: easy to use, portable and resource efficient operating system for networked embedded devices," in *Proceedings of the 8th ACM Conference on Embedded Networked Sensor Systems*, ser. SenSys '10. New York, NY, USA: ACM, 2010, pp. 427–428. [Online]. Available: <http://doi.acm.org/10.1145/1869983.1870057>

Microstrip Patch Antenna Based on Photonic Crystal Substrate with Heterostructures at Terahertz Frequency

Farhad alizadeh*, Alireza maleki javan**, and Manoochehr kamyab hesari***

* Farhadiou@yahoo.com , ** amalekijavan@eedt.kntu.ac.ir, and *** kamyab@eedt.kntu.ac.ir
communication Engineering Dep. ,Islamic azad university of Tehran south, Tehran, Iran
Tell: (+9821) 33434278

Abstract—In this paper, a patch antenna with an PBG (photonic band-gap) structure was studied by FDTD method .The results indicate that, comparing to the conventional patch antennas, the expansion of working frequency band of the new patch antenna can be realized and its radiation efficiency also can be improved notably with the influence of Photonic crystal substrate.

KeyWords- 2-D Photonic crystal, metamaterial, microstrip patch antenna, radiation efficiency, terahertz spectrum, directivity.

1. Introduction

The concept of “photonic crystal (PC)” was firstly put forward by E. Yablonovitch [1] and S. John [2] in the year of 1987. The introduced dielectric periodicity in PCs can modulate the propagation of electromagnetic waves due to the multiple Bragg scatterings and results in the relationship of frequency versus wave-vector characterized by photonic band structure. An electromagnetic wave with frequency falling in the “forbidden band” or “stop band”, is forbidden from propagating in any direction in the PC [3,4]. A non-transmission frequency range will be expected with the complete reflection to those electromagnetic waves falling in the non-transmission range. It has also been known that the PBG is determined by the topology, dielectric constant, and the connectivity [5,6]. Semiconductor heterostructures have revolutionized optoelectronics and high-speed electronics through their ability to confine electrons of precise kinetic energies in specific devices areas. Photonic crystal heterostructures provide a similar amount of control over the wavelength and localization of light in photonic crystals [7-10]. This is because the strong light-matter interaction in heterostructures of PCs can result in their different positions of the forbidden bands in energy coordination for the different heterogeneity structures of PCs. This also

renders an enlargement of non-transmission frequency band in PBG heterostructures in comparison to each single PBG in the heterostructures.

Nowadays, photonic crystals have been found many applications in microwave circuits, antennas [11-14], etc. Photonic crystal patch antennas or PBG patch antennas are very useful because the photonic crystals can reduce the substrate absorption compared with conventional patch antennas without PBG substrate due to the existence of non-transmission bands of PCs. The enlargement of non-transmission frequency band in photonic crystal heterostructures will further reduces the substrate absorption and improve the radiation efficiency of antenna.

In the paper, the comparison is carried out to the performance of (1) a conventional patch antenna without PBG substrate and (2) a patch antenna with PBG heterostructure substrate by employing the finite difference time domain (FDTD) method. The results show that, comparing to the conventional patch antennas, the working frequency band of the photonic crystal patch antenna with heterostructures can be expanded due to the enlargement of non-transmission frequency band. And its radiation efficiency as well as its return loss can also be improved notably. The physical mechanism lies in the PBG which suppress the surface waves propagating along the surface of the substrate and reflect most of electromagnetic wave energy radiated to the substrate significantly. Due to such advantages, the use of photonic crystal patch antennas will be extended in many areas, such as, mobile communications, satellite communications as well as communications for aeronautics and astronautics.

THz time-domain spectroscopy uses coherent pulses of electromagnetic radiation to obtain information about the frequency range between 0.1 and 10 THz. The electromagnetic pulses have sub-picosecond width and a

field strength on the order of 10-100 V/cm, depending on emitter geometry.

2. CALCULATION MODEL

The FDTD method [15-18] is frequently employed for calculating the PBG patch antennas due to its advantages in comparison with other algorithms used for the same purposes. Its fundamental principle is that Maxwell's equations are expressed as scalar equations of electric and magnetic field components in Cartesian coordinates first, and then replaces differential quotient with difference quotient with second-order precision. Each photonic crystal cell is meshed with the method proposed by Yee [15], here Δx and Δy are the lattice space increments in the x-and y-coordinate directions, respectively, and Δt is the time increment. By discretizing partial differential Maxwell's equations in the space and time domains in Yell cell, one can obtain the following FDTD approximation as representative relations with respect to TE modes,

$$E_x^{n+1}(i,j) = E_x^n(i,j) + \frac{H_z^{n+\frac{1}{2}}(i,j+\frac{1}{2}) - H_z^{n+\frac{1}{2}}(i,j-\frac{1}{2})}{\Delta y} \cdot \frac{\Delta t}{\epsilon(i,j)} \quad (1)$$

$$E_y^{n+1}(i,j) = E_y^n(i,j) + \frac{H_z^{n+\frac{1}{2}}(i+\frac{1}{2},j) - H_z^{n+\frac{1}{2}}(i-\frac{1}{2},j)}{\Delta x} \cdot \frac{\Delta t}{\epsilon(i,j)} \quad (2)$$

$$H_z^{n+1}(i,j) = H_z^{n-\frac{1}{2}}(i,j) + \frac{E_x^n(i,j+\frac{1}{2}) - E_x^n(i,j-\frac{1}{2})}{\Delta y} \cdot \frac{\Delta t}{\mu} - \frac{E_x^n(i+\frac{1}{2},j) - E_x^n(i-\frac{1}{2},j)}{\Delta x} \cdot \frac{\Delta t}{\mu} \quad (3)$$

To ensure steady results in iteration constringency, Δx , Δy , Δt must satisfy the steady condition [15],

$$\Delta t \leq \frac{1}{c \sqrt{(\Delta x)^{-2} + (\Delta y)^{-2}}}$$

an approximate formula for the radiated fields from a rectangular patch antenna can be derived. For this analysis, the E-field is assumed to be uniformly distributed along the patch with amplitude $E_{ax} = E_0$. Based on this theory, a uniform source radiating in the yz-plane can be evaluated using [23]

$$P_x = \iint_{sa} E_{ax}(y',x') e^{j\beta(y' \sin\theta \cos\varphi + z' \cos\theta)} dy' dz' \quad (4)$$

$$E_\theta = j\beta \frac{e^{-j\beta r}}{2\pi r} (P_x \cos\varphi) \quad (5)$$

$$E_\varphi = j\beta \frac{e^{-j\beta r}}{2\pi r} \cos\theta (-P_x \sin\varphi) \quad (6)$$

Equation (4) reduces to

$$P_x = E_0 \frac{\sin(Y)}{Y} \frac{\sin(Z)}{Z} \quad (7)$$

$$Y = \frac{\beta W}{2} \sin\theta \sin\varphi \quad (8)$$

$$Z = \frac{\beta h}{2} \cos\theta \quad (9)$$

For very thin substrates ($bh \ll 1$) and as the limit of $Z \rightarrow 0$, equation (7) reduces to

$$P_x = E_0 \frac{\sin(Y)}{Y} \quad (10)$$

This expression is valid for only a single radiating slot. The expression when both slots radiate simultaneous is calculated using array theory. The normalized array factor for two elements, of the same magnitude and phase, separated by a distance L along the x axis is

$$AF(\theta, \varphi) = \cos\left(\frac{\beta l}{2} \sin\theta \cos\varphi\right) \quad (11)$$

The far-field expression for a patch are evaluated from equations (5) and (6) to produce

$$E_\theta = E_0 \cos\varphi (f(\theta, \varphi)) \quad (12)$$

$$E_\varphi = -E_0 \cos\theta \sin\varphi (f(\theta, \varphi)) \quad (13)$$

$$f(\theta, \varphi) = \frac{P_x}{E_0} AF(\theta, \varphi) = \frac{\sin\left(\frac{\beta W}{2} \sin\theta \sin\varphi\right)}{\frac{\beta W}{2} \sin\theta \sin\varphi} \cdot \cos\left(\frac{\beta W}{2} \sin\theta \cos\varphi\right) \quad (14)$$

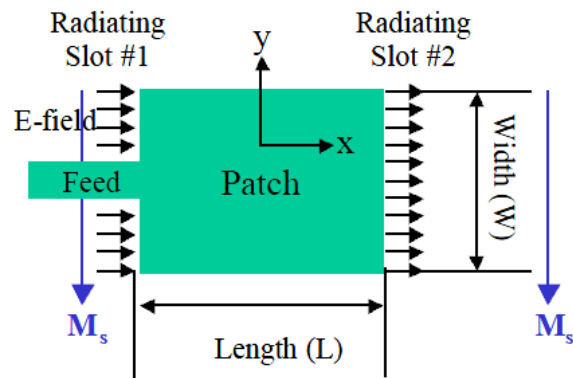


Fig.1 Top view showing the fringing electric fields that is responsible for radiation along with the equivalent magnetic surface (M_s) currents.

The cavity model provides a description of how the fields radiating from the surface of the patch antenna.

Controlling the radiation characteristics of the patch with photonic crystals is the main principle in this paper.

3. ANTENNA CONFIGURATION

Fig. 1 shows the geometrical configuration of proposed antenna. In this configuration, host material of the substrate is Gallium arsenide ($\epsilon_r=12.9$) of thickness $50\mu m$. In the host material, The radius of the cylinder holes are all $13\mu m$, while the side lengths of the square holes are all $18\mu m$, and the distance between two neighboring cylinder or square centers is $100\mu m$. The substrate dimension is $500\mu m \times 500\mu m$. The size of the radiating patch on this substrate has been calculated equal to $60\mu m \times 60\mu m$. The length of the patch is equal to free space half wavelength at 680 GHz. To match the impedance of resonator to the port, gap coupled feeding technique has been employed. In this way, the length and width of feed line has been selected equal to $118.02\mu m$ and $20\mu m$, respectively.

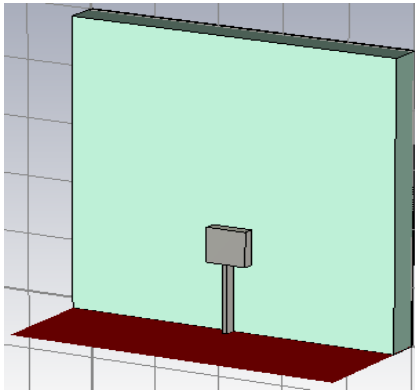


Fig.2 Structural parameters of conventional patch antenna (simulated by CST).

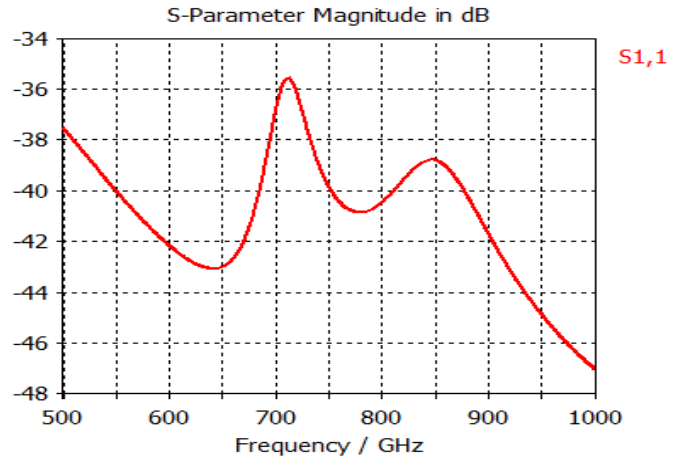


Fig.5 Return loss S11 from the conventional patch antenna (simulated by CST).

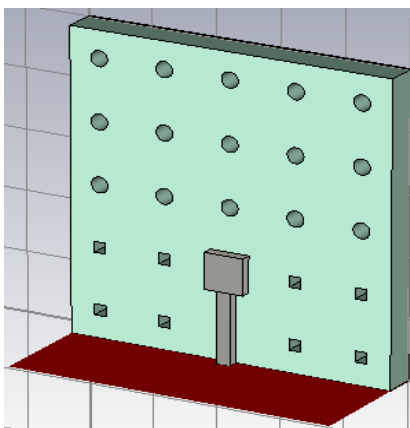


Fig.3 Structural parameters of PBG patch antenna (simulated by CST).

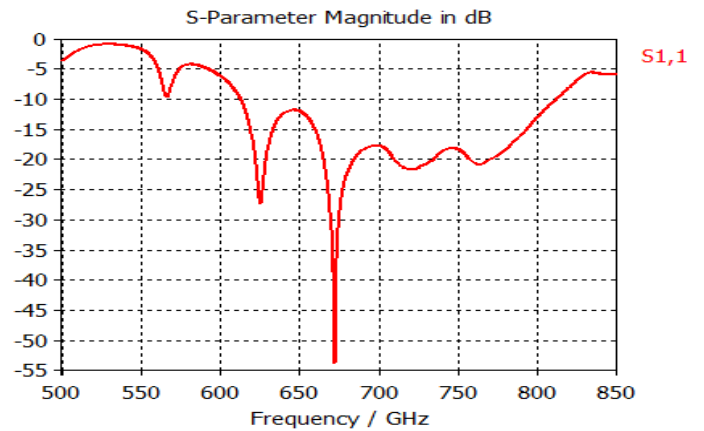


Fig.6 Return loss S11 from the PBG patch antenna (simulated by CST).

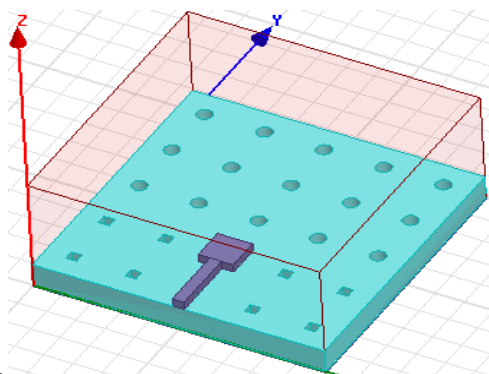


Fig.4 Structural parameters of PBG patch antenna (simulated by HFSS 13).

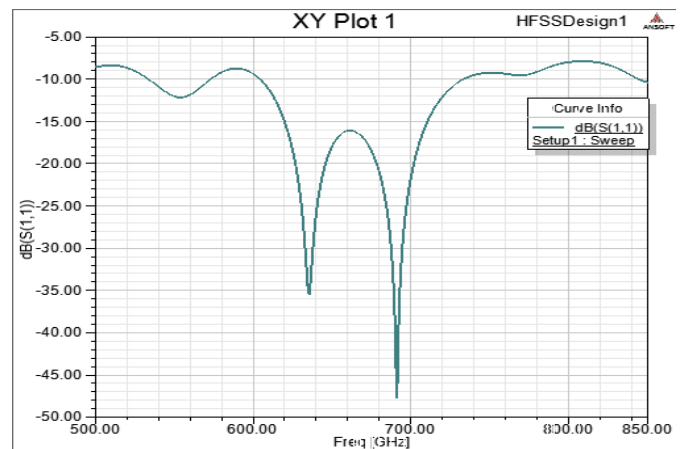


Fig.7 Return loss S11 from the PBG patch antenna (simulated by HFSS 13).

4. SIMULATION RESULTS AND ANALYSIS

To simulate these two patch antennas with structure parameters shown in Fig.2, Fig.3 and Fig.4, we solve Maxwell's equations by the method of FDTD [20]. Return loss (S11) are shown in Fig.5-6-7, respectively.

Comparing Fig.5 with Fig.6-7, one can find out that the return loss (S11) of the patch antenna with PBG substrate is improved notably, with minimum return loss about -54dB at the frequency of 672GHz and -28dB at the frequency of 625.5GHz corresponding to the resonant frequencies of each PC in the substrate has been simulated by using CST Microwave Studio, and minimum return loss about -48dB at the frequency of 684GHz and -36dB at the frequency of 628GHz corresponding to the resonant frequencies of each PC in the substrate by using HFSS 13, about 13dB lower than of the conventional patch antenna. It indicates at the same time that the working frequency band patch antenna can be extended. As lower return loss and higher gain from the PBG patch antenna that were mentioned can be stated easily by the theoretical point of view. A band gap is formed by adding the PBG structure in the substrate, and EM waves with frequencies falling in the band gap will be suppressed, namely, they can not be transmitted in any direction in the substrate.

At the same time, according to the antenna radiation patterns, we can find that the conventional antenna's maximum gain is 6.44dB, while that of the PBG one is 9.234dB. High gain of the PBG patch antenna results from the effect of electromagnetic waves being highly localized by the aberrance between the PCs, that is to say, the electromagnetic field has enhanced strongly at the interface. It suggests that the PBG structure can improve patch antennas' gain obviously. Therefore, the absorption of EM waves by the substrate is reduced, and their energy is reflected back into the free space, so its return loss and gain are improved greatly.

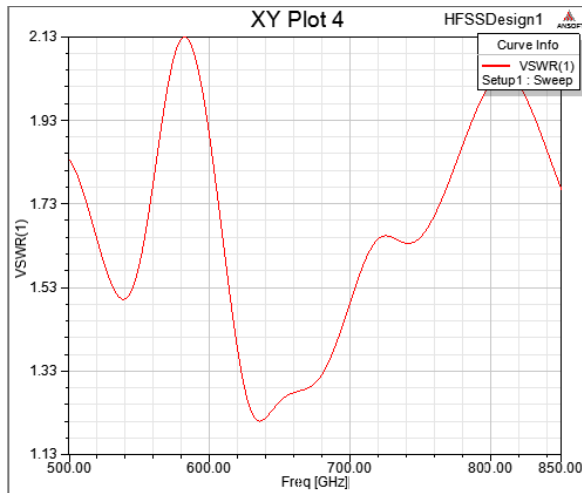


Fig.8 VSWR from the conventional patch antenna.

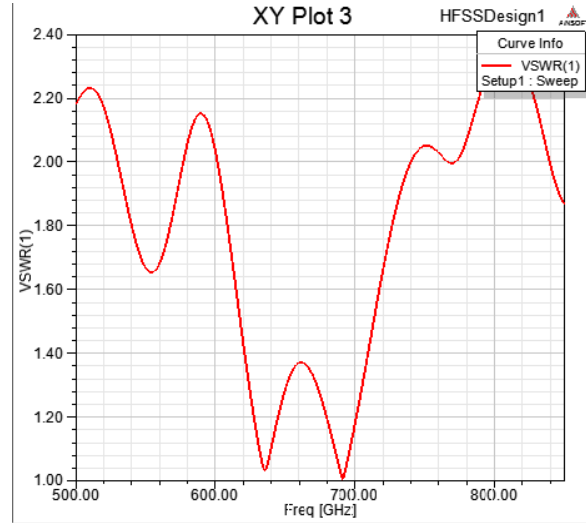


Fig.9 VSWR from the PBG patch antenna.

Comparing Fig.8 with Fig.9 For the VSWR (voltage standing wave ratio), it is 1.00126 for PBG patch antenna, which is very close to the ideal value 1.0, while it is 1.28 for the conventional patch antenna. Main characteristic parameters of the patch antennas drawn by the simulation are listed in Table 1.

TABLE I
MAIN CHARACTERISTIC PARAMETERS OF THE PATCH ANTENNAS

	Return loss	Directivity	Maximum gain	VSWR
PBG patch antenna	-54dB@ 742GHz	9.128dB _i	9.234dB	1.00126
Conventional patch antenna	-43dB@ 650GHz	6.182dB _i	6.44dB	1.28

5. CONCLUSION

In this paper, we have designed a microstrip patch antenna on a 2-D photonic crystal substrate with Heterostructures at Terahertz Frequency. Obviously, the electric field energy from surface waves absorbed by conventional antenna is much more than that absorbed by the PBG patch antenna. It shows that surface waves transmitted along the substrate can be suppressed by PBG structure. The gain of this antenna is 9.234dB and directivity is 9.128 dB_i at two resonant as well as intended frequencies. The high gain of this antenna would enable to combat the losses at terahertz frequency. the structure has been simulated by two different simulators CST Microwave Studio and Ansoft HFSS 13.

REFERENCES

- [1] Yablonovitch E, "Inhibited spontaneous emission in solid state physics and electronics," [J]. Phys. Rev. Lett., 1987, 58(20), pp.2059-2062.
- [2] John S, "Strong localization of photon in certain disordered dielectric super lattice," [J].Phys. Rev. Lett.,1987,58(23), pp. 2486 -2489.
- [3] Soref R, "The Achievements and Challenges of Silicon Photonics", [J]. Advances in OptoElectronics, 2008, article ID 472305.

- [4] Seigneur H P, Weed M, Leuenberger M N, and Schoenfeld W V, "Controlled On-Chip Single-Photon Transfer Using Photonic Crystal Coupled-Cavity Waveguides", [J]. *Advances in OptoElectronics*, 2011, article ID 893086.
- [5] Yablonovitch E, Gmitter T J, and Leung K M, "Photonic Band Structure: The Face-Centered-Cubic Case Employing Nonspherical Atoms", [J]. *Phys. Rev. Lett.*, Oct. 1991, pp.2295-2299.
- [6] Kitzerow H-S, Matthias H, Schweizer S L, H M van Driel, and Wehrspohn R B, "Tuning of the optical properties in photonic crystals made of macroporous silicon", [J]. *Advances in Optical Technologies*, 2008, article ID 780784.
- [7] E.Yablonovitch, E Kapon, T.J.Gmitter, C.P. Yun, R. Bhat, "Double heterostructure GaAs/AlGaAs thin film diode lasers on glass substrates," *IEEE Trans. Photonics Technology Letters*, Feb. 1989, pp.41-42.
- [8] Lin L L, Li Z Y, "Interface states in photonic crystal heterostructures" *Phys Rev B*, 2001, 63(3), pp. 033310-1-4.
- [9] Zhou Y S, Gu B Y, Wang F H, "Guided modes in photonic crystal heterostructure composed of rotating non-circular air cylinders in two-dimensional lattices," *J Phys: Condens. Matter*, 2003,15(10), pp.4109-4118.
- [10] Zhou Y S, Gu B Y, Wang F H, "Photonic band-gap structures and guide modes in two-dimensional magnetic photonic crystal heterostructures," *Eur Phy J B*, 2004,37(4) pp. 293-299.
- [11] Yasushi Horii, Makoto Tsutsumi, "Harmonic control photonic bandgap on microstrip patch antenna," *IEEE Microwave Guided Wave Lett.*, 1999, 9(1), pp.13-15.
- [12] Joseph S. Colburn, Yahya Rahmat-Samii, "Patch antennas on externally perforated high dielectric constant substrates," *IEEE Trans. on Antennas Propagation*, 1999, 47(12), pp.1785-1794.
- [13] Shen Ting-gen, Zhou Yue-qun, Li Zheng-hua, Ji Pei-lai, Sun Jin, Yu Feng-chao, "Influence of a novel periodic structure on patch antenna," *Appl Phys A* (2009) 96: 789–792.
- [14] Zhou Yue-Qun, Yu Feng-Chao, Shen Ting-Gen, Ji Pei-Lai, Ge Jun, Gen Jian-Feng, Len Jing, "Investigation of Patch Antennas Based on Embedded Multiple PBG Structure," *IEEE PHOTONICS TECHNOLOGY LETTERS*, (2008) 20:1685-1687.
- [15] K. S. Yee, "Numerical solution of initial boundary value problems involving Maxwell's equations in isotropic media," *IEEE Trans. Antennas and Propagation*, 1966.14(3), pp. 302-308.
- [16] Qiu M, He S L, "Numerical method for computing defect modes in two-dimension photonic crystals with dielectric of metallic inclusions," [J] *Phys Rev B*, 2000,61(19), pp.12871-12876.
- [17] Dong Xiaoting, X. S. Rao, Y. B. Gan, B. Guo and W. Y. Yin, "Perfectly matched layer absorbing boundary condition for left handed materials," *IEEE Microwave and Wireless Components letters*, 2004,14(6), pp.301-303.
- [18] Hamidreza Azarina, Ahad Tavakoli, "Finite difference time domain analysis of a photonic crystal substrate patch antenna," *Physica B* 2005, 370, pp. 223–227.
- [19] Fang Zhao, Libo Yuan, "Interface guided modes in phononic crystal heterostructures," *J. Phys. D: Appl. Phys.* 39 (2006) 4783–4787.
- [20] Okada N, Cole J B, Yamada S, Ogawa K, and Katayama Y, "Nonstandard FDTD simulation-based design of CROW wavelength splitters", [J]. *Advances in Optical Technologies*, 2011, article ID 265702.
- [21] E.R. Brown, C.D. Parker and E. Yablonovitch, "Radiation properties of a planer antenna on a photonic-crystal substrate," *J. Opt. Soc. Am. B*, Vol 10, No.2, pp. 404-407, February 1993.
- [22] Zhenghua Li, Yan Ling Xue, Tinggen Shen, "Investigation of Patch Antenna Based on Photonic Band-gap Substrate with Heterostructures", *Hindawi Publishing Corporation, Journal of Antennas and Propagation*, 2012, article ID 151603.
- [23] Stutzman and Thiele, "*Antenna Theory and Design*", 2nd ed: John Wiley & Sons, Inc., 1998.
- [24] Balanis, "*Antenna Theory Analysis and Design*", 2nd ed: John Wiley & Sons, Inc., 1997.

Designing Yagi-Uda Antenna Fed by Microstrip Line and Simulated by HFSS

Hassan Karbalaee
Shahed University

Tehran – Iran
H79_karbalaee@yahoo.com

Mohammad Reza Salehifar
Islamic Azad University

Science and Research Branch
Tehran – Iran
Mr_salehifar@yahoo.com

Saeed Soleimany
Islamic Azad University

Qazvin Branch
Qazvin - Iran
Saeed.soleimany@gmail.com

Abstract- In this article, we have proposed a novel Yagi antenna that has both compactness of resonant antenna and broadband characteristics of traveling-wave radiators. It is fabricated on dielectric substrate with a microstrip (MS) feed. The top layer consists a microstrip feed, a broadband microstrip-to-coplanar stripline (CPS) transition and two dipole-elements, one of them which is driver element is fed by CPS, and the second is director. The metal underside is a microstrip ground which serves as a reflector element and cancels using reflector dipole. This antenna is constructed on low cost substrate with 1.56mm thickness and $\epsilon_r = 4.12$. The simulated bandwidth is about 35% and its gain is more than 5dB.

Keywords: Yagi-Uda antenna, HFSS, microstrip

I. INTRODUCTION

Yagi-Uda first recalls television antenna installed on the roofs. An example of three elements antenna is shown in Fig.1. The optimal spacing in order to maximum directivity (about 15dB) between driver and reflector is about 0.15λ and the rate between driver and director is about 0.25λ . Generally, the reflector is 5% or more longer than resonant length of driver, similarly the director is shorter. In General, for antenna with three or more elements, the driver with a length less than $\lambda/2$ ($0.45\lambda - 0.49\lambda$) will resonate while the directors length is about ($0.4\lambda - 0.45\lambda$). Typically the distance between directors is ($0.3\lambda - 0.4\lambda$) which is not necessary be similar for optimal design [1].

II. YAGI ANTENNA BASED ON MICROSTRIP AND COPLANAR FEEDING

In order to increase the distance of communications, reducing interferences coming from other wireless system also to provide wider bandwidth, the use of printed dipole antenna is highly recommended [2]. Our antenna schematic which is shown in Fig.2 is designed for performing in UHF band with about 500~800 MHz bandwidth. Microstrip feeding has this benefit that the dielectric used in its construction is virtually a mechanical protective.

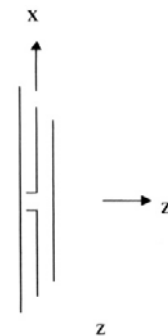


Fig.1 Three elements Yagi Uda antenna

A printed dipole (driver) is used in order to create TE_0 mode surface wave with the least unintended TM_0 mode that is effective in Cross-Polarization level determining. In this plan, the director in addition directs the waves of antenna towards the endfire direction, simultaneous is involved as a impedance matching element and the groundplane of the microstrip line on the feeding port of CPS acts as a reflector. Entirely these cases lead to compression circuit and compatibility with each MMIC circuit based on microstrip [3].

First, feeding lines and antenna is designed, simulated and optimized with HFSS software and finally a practical model is made.

III. MICROSTRIP LINE AND POWER DIVIDER DESIGN

Microstrip line is designed with 50Ω impedance. There are specific software to calculate the line parameters such as TX Line, that each calculation methods differ from each other. Also by referring to microstrip line equations [4], we can separately achieve these line parameters :

$$W=2.94mm, \epsilon_{re} = 3.13, \lambda_g = 84.78mm$$

In order to convert MS to CPS for feeding antenna is required transmitted power from 50Ω line divide between two 50Ω lines. For this reason we need a impedance transformer, is shown in Fig.3 .

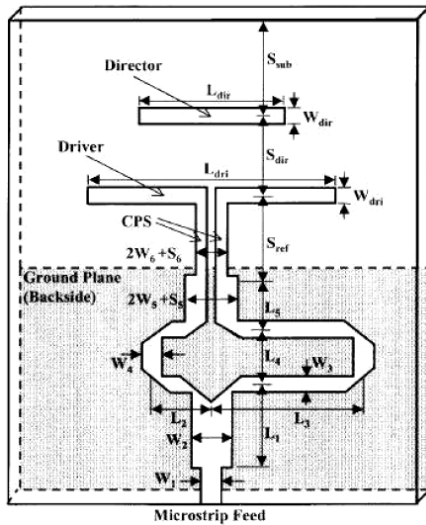


Fig.2 Microstrip Yagi antenna schematic

This is a quarter wave converter for matching 25Ω to 50Ω . By referring to converters formulas [5]:

$$Z_0 = \sqrt{Z_{in} \times Z_L}$$

So transformer characteristic impedance with $\lambda_g/4$ length equal to 35.35Ω . By referring to [4]:

$$W = 5.08mm$$

Of course with connecting three 50Ω lines to this transformer, its size needs to resize. In this process, we assumed width and length transformer as variable elements and set some parameters of simulation in HFSS as follows:

(width) $W_t = \text{variable}$

(length) $L_t = \text{variable}$

Solution frequency = 2.1GHz

Maximum Number of Passes = 15

Minimum Number of Passes = 3

Minimum Number of Converged Passes = 2

$$\Delta s = 0.02$$

New dimensions of simulation are:

$$W_t = 5.2mm$$

$$L_t = 20.2mm$$

In Fig.4 S parameter curve is drawn from simulation which shows the lowest return loss in 2GHz.

IV. 180° DELAY LINE

By using the balun phase shifter to generate a 180° phase difference between the coupled microstrip lines at the working frequency, the correct excitation to the antenna is provided [2]. It is done with selecting proper length for L_2, L_3 in Fig.2 such that

$$L_3 - L_2 = \lambda_g/4$$

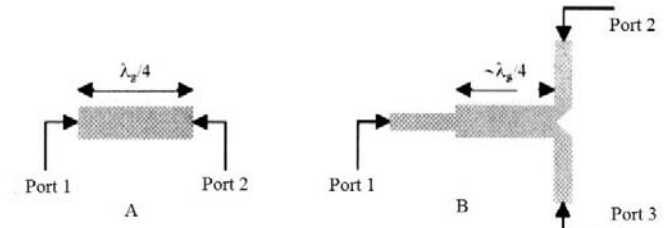


Fig.3 A. The quarter wave converter with 2 ports
B. The quarter wave converter with 3 ports

This cause odd mode as a dominant mode in coupled microstrip line therefore feeding for CPS will be balanced. In order to the least effect on the phase and amplitude of the signal in 90° bends, we use mitred bends. The distance between cut away point and outer corner of the un-mitred bend is:

$$y = 0.42mm$$

This value is obtained by using formulas presented in reference [6].

V. COPLANAR STRIP LINE AND COUPLED MICROSTRIP

Since CPS line dose not accept even mode, playing as open circuit for even mode of coupled microstrip line and allows us to negate unintended excited even mode in couple line. 100Ω line with $W = 5.17mm$ and $L = 0.6mm$ is calculated by formula No.7.

The connected coupled MS to CPS line is designed for 50Ω characteristic impedance. Based on [7] formula, will be obtained:

$$W = 2.94mm, S = 0.6mm$$

Coupled line width is considered as equal as 50Ω MS line width because of having the same mitred bends. Fig.5 is equivalent circuit of odd mode excitation demonstrator.

VI. RADIATOR ELEMENT

Initial dimensions antenna are chosen such as normal yagi mentioned in introduction. Main part of antenna, reflector, which located on back of the board as a finite ground plane, driver feeding by CPS and director are designed in HFSS. For a 100Ω input impedance, the best response for much bandwidth will be obtained with these dimensions:

$$W_{dri} = 5.3mm, L_{dri} = 67.7mm$$

$$W_{dir} = 5.3mm, L_{dri} = 33.9mm$$

$$S_{dir} = 19.5mm, S_{sub} = 19mm, S_{ref} = 35.2mm$$

VII. RADIATOR ELEMENT

With connecting the feeding part to radiator elements Fig.6 will be obtained. Since this step is the last one in this simulation and need more accuracy, the type of solution in

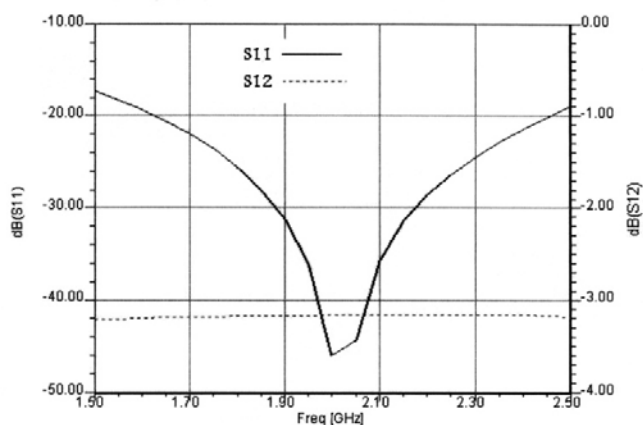


Fig4. Simulated S parameter of power divider

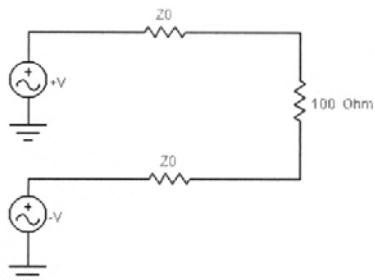


Fig5. Equivalent circuit of odd mode excitation

HFSS is chosen DRIVEN MODE and the antenna will be located in an air box that really makes radiation boundary; while the distance between cubic surfaces and antenna is more than λ . The input port is LUMPED PORT type with 50Ω characteristic impedance and also INTEGRATION LINE is drawn from MS line to ground. SOLUTION FREQUENCY is considered 2.1GHz for better accuracy. Other parameters are the same as part III.

After the latest optimization on all of dimensions specially dimensions of antenna and CPS line to achieve the lowest return loss, the following results for antenna dimensions is obtained that its details are shown in Table 1.

VIII. SIMULATION AND PRACTICAL TEST RESULTS

According to the simulation, the antenna has a bandwidth ($S_{11} < -10dB$) about 37% that is shown in Fig.7 . The input impedance in Fig.8 shows the antenna in this bandwidth has a good matching.

The antenna patterns in $\varphi = 90^\circ, \varphi = 0^\circ$ planes in Fig.9 and also the obtained gain from simulation in bandwidth range in Fig.10 are shown.

As those Figures show, at the center frequency (2GHz) of bandwidth , gain is 5.4dB and Front to back Ratio is 13dB. Also Fig11,12 illustrate in HPBW range, Cross Polarization Surface is lower than -24dB and this is so desirable.

Table 1. Final antenna dimensions based on mm

W_1	2.94	S_5	0.6
W_2	5.2	W_6	5.3
L_1	22.09	S_6	0.6
L_2	5.55	L_{dri}	67.7
L_3	26.75	W_{dri}	5.3
W_3	2.94	L_{dir}	33.9
W_4	2.94	W_{dir}	5.3
L_4	9.78	S_{ref}	35.2
L_5	9.09	S_{dir}	19.5
W_5	2.94	S_{sub}	19

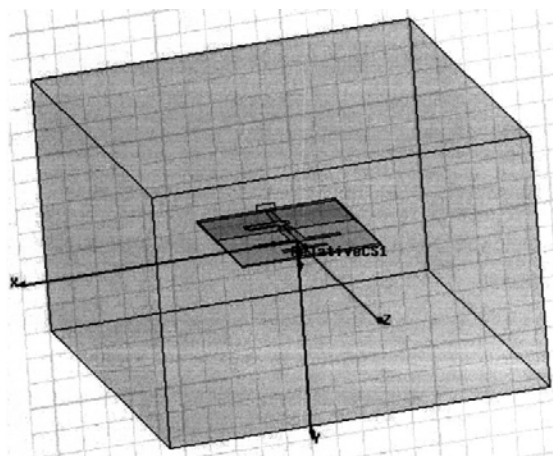


Fig6. Equivalent circuit of odd mode excitation

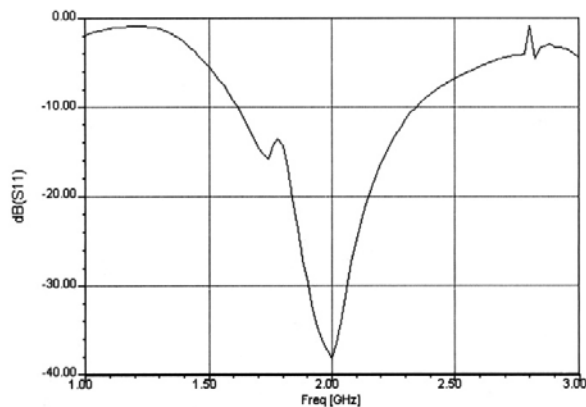


Fig7. Simulated Return Loss of Antenna

According to these results, an example of antenna was made that its measurement results in Fig13,14 shows a little difference between simulation and reality; the following reasons can be outlined for this difference:

- Using non-professional version of HFSS
- Style of connecting the connector and solder

IX. CONCLUSION

This configuration of printed Yagi-Uda antenna is designed and demonstrated. The optimal dimensions of the antenna are obtained by HFSS simulation. Even with thicker board by greater ϵ_r , the dimensions will be shorter. With good characteristics such as wide bandwidth and suitable FTBR, Cross Polarization and gain, it should find wide applications in wireless systems such as WLANs or GSM. Despite using only one director, it has more gain and a wider bandwidth compared to similar designed antennas. For more gain, adding parasitic elements to the basic antenna structure or using antenna arrays is recommended. If we use a printed dipole with arms in both sides of dielectric substrate in opposite directions, and feeding by a transmission line, placed on both sides of the substrate[7][8], the balun phase shifter is deleted and final antenna size is reduced. Also the gain is increased but the bandwidth is slightly decreased.

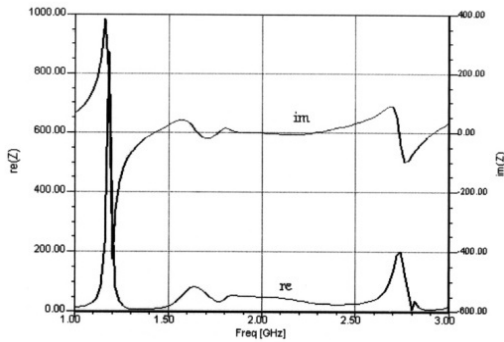


Fig8. Simulated Input Impedance of Antenna

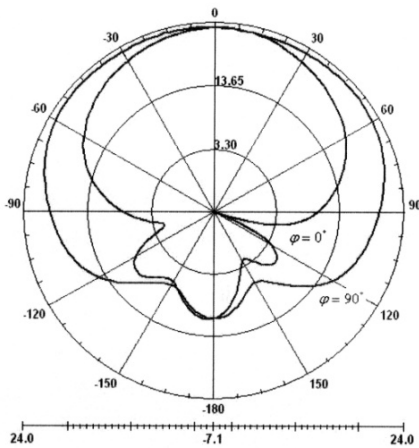


Fig9. Simulated Radiation Pattern

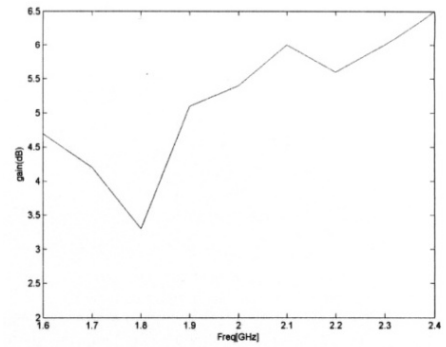


Fig10. Simulated Gain in Bandwidth Range

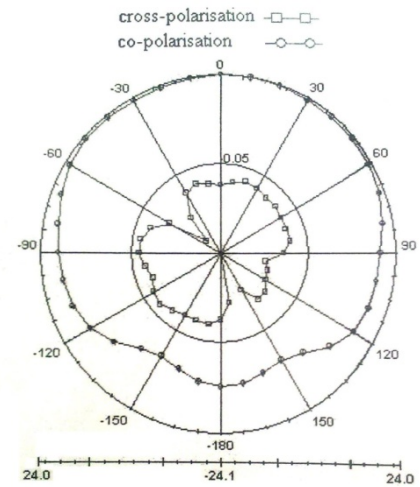


Fig11. Simulated Radiation Pattern in H plane

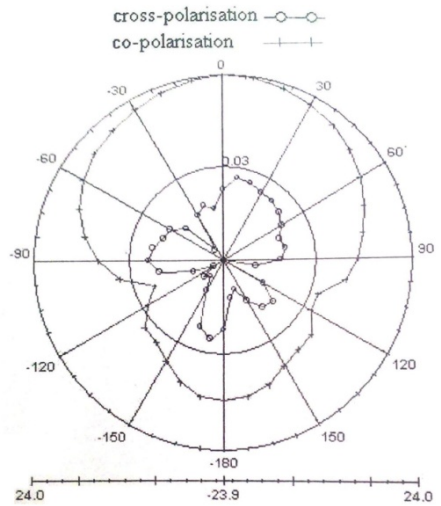


Fig12. Simulated Radiation Pattern in E plane

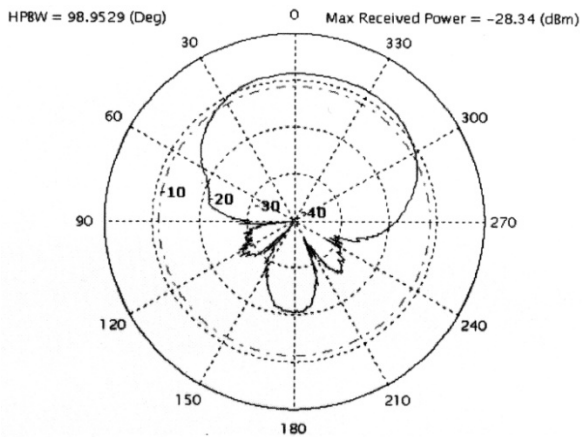


Fig13. Measured Radiation Pattern for $\varphi = 0^\circ$

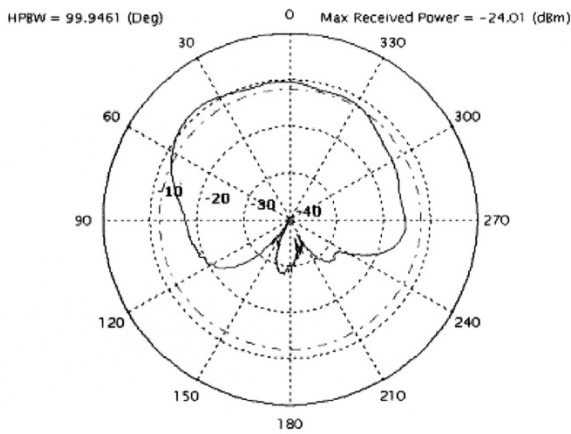


Fig14. Measured Radiation Pattern for $\varphi = 90^\circ$

X. REFERENCES

- [1] Warren L. Stutzman, Gary A. Tiele, "Antenna Theory and Design", John Wiley & Sons, Inc, 1997.
- [2] E. A' vila-Navarro, J. A. Carrasco, and C. Reig, "Printed dipole antennas for personal communication systems," IETE Technical Review, vol. 27, no. 4, pp. 286-292, 2010.
- [3] W.R.Deal, Noriaki Kaneda, James Sor, Yongxi Qian and T.Itoh, "A New Quasi Yagi Antenna for Planar Active Antenna", *IEEE Transactions on Microwave Theory and Techniques*, vol. 48, no.6, June 2000.
- [4] Rajeh Mongia, Inder Bahi, Prakash Bhartia, "RF and Microwave Coupled-Line Circuits", Artech House, 1999.
- [5] Peter A.Rizzi, "Microwave Engineering", Prentice Hall, 1987.
- [6] E.H.Fooks, R.A.Zakarevicius, "Microwave Engineering Using Microstrip Circuits", Prentice Hall, 1989.
- [7] A.cheldavi, G. Rezai Rad, "Introduction to Electromagnetic Compatibility", Iran University of Science & Technology, 2004.
- [8] E. A' vila-Navarro, C. Reig, "Directive Microstrip Antennas for Specific Below -2.45 GHz Applications", Hindawi Publishing Corporation International Journal of Antennas and Propagation, Volume 2012

Multipurpose Smart SIM Card Based on Mobile Database and Location Dependent Query

Hamid-Reza Firoozy-Najafabadi
Department of Computer Engineering
Science and Research Branch, Islamic Azad University
Tabriz, Iran
hr.firoozy@iauotash.ac.ir

Mohammad-Reza Feizi-Derakhshi
Department of Computer Science
University of Tabriz
Tabriz, Iran
mfeizi@tabrizu.ac.ir

Abstract— According to increasing development of technology and in order to approaching electronic government, most citizen services are presented electronically using smart electronic cards. Regarding this fact, people have several smart cards which are increasing in number everyday. Despite of so many advantages, these cards have various disadvantages such as multiplicity, troublesome carrying, unavailability of card readers in many places, waiting in the queues of ATMs, etc. In this paper we will present a new approach named multipurpose smart SIM card in order to solve these problems. Proposed SIM card, acts based on mobile database architecture that we will discuss about this architecture and location dependent queries processing in it. Then we will study the problem of finding the nearest and most unoccupied ATM as a case study and at last we will propose a method to authenticate users in this system.

Keywords- smart SIM card; mobile computing; mobile database; location dependent query; authentication

I. INTRODUCTION

Today's world is a technological world and its dominant atmosphere induces human toward mechanization and facilitation. In this case and across growing procedure of technology, concept of electronic card has been established. Such cards help us doing our tasks timely and reliably.

Nowadays smart cards are used as bank card, telephone card, subway credit card, fuel card, buying card, health card, etc. [10]. And the number of these cards is increased everyday and some other new electronic cards like health electronic card, passport electronic card, driving license card and so many other electronic cards will be required. So one day in a future we should carry so many electronic cards every time we want to go out. Another problem is the limitation and low availability of card readers. It means that when you are at home, in your car, out of city or other places where there are not any card readers or ATMs, in this case you are not able to use your cards.

The solution of these problems is an integration of smart cards that some countries have attempted to produce a national

smart card [9, 10] in order to integrate a variety of smart card information on one card to fix the problems. The national smart card is a plastic card about the size of a credit card, with an embedded micro-chip to deliver one or more intelligent capabilities. The national smartcard framework is one of a number of frameworks and strategies developed to support interoperable whole-of-government business applications [10].

In this paper we propose a method to this integration and that is multipurpose smart SIM card. A subscriber identity module or SIM is a smart card that is designed to fit into mobile phone. It provides the identification of a user to a network, allowing him or her to access such services as telephony, email, internet and text messaging. The SIM card contains a microcomputer as well as a certain amount of memory named Random Access Memory (RAM) to process commands and Electronically Erasable Programmable Read Only Memory (EEPROM) to store user files. The SIM card also contains an amount of Read Only Memory (ROM) which stores the cards operating system. When the SIM card is activated the microcomputer loads the operating system from ROM into the RAM of the card and processes commands as requested by the mobile equipment (ME) or card access device (CAD). Naturally architecture and foundation network of this plan should follow a particular standard that we will discuss about this issue further.

After this introduction, paper is organized as follows: section 2 provides an overview of mobile database architecture and in section 3 queries processing in the mobile database is discussed. Section 4 presents the proposed method to integrate smart cards and section 5 presents a case study to illustrate a practical application of the proposed method. Finally conclusion and future work is presented.

II. THE ARCHITECTURE OF MOBILE DATABASE

Proposed SIM card database should be implemented based on mobile database architecture. A mobile database is a kind of distributed database that supports mobile computing [4]. It means that its database is distributed on some fixed and mobile parts so that across movement of users and location changing of their mobile phones, wireless network has capability of processing of transactions and data management.

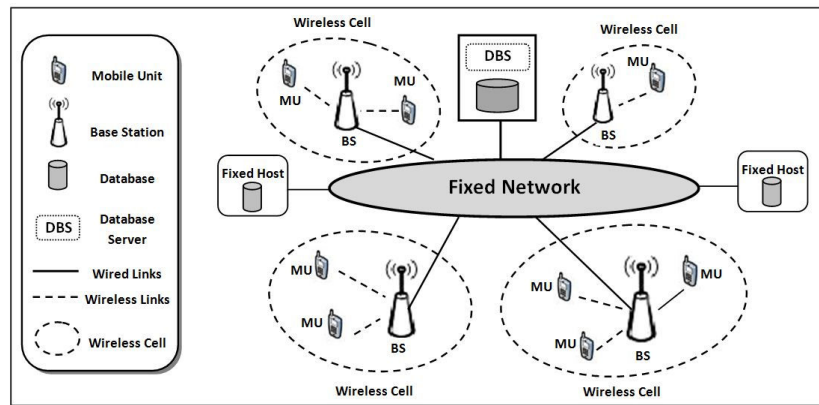


Figure 1. Mobile database architecture

Figure 1 shows mobile database architecture which includes:

- *Fixed Network (FN)*: Is a conventional wired network which interconnects some other stationary parts [1, 3].
- *Fixed Host (FH)*: This part has a fixed place in network and it does not have wireless interface. Therefore mobile units can't connect to it directly [1, 5].
- *Cell*: Network covered area is divided into some cells that the size of these cells depends on the power of base station [1, 5]. Mobile units move alternatively among different cells and have communication with different support stations at different times.
- *Mobile Unit (MU) or Mobile Host (MH)*: These are devices that are moving and they have the capability of connecting to fixed network through wireless connection. Users send their transaction requests through MU and receive results.
- *Mobile Support Station (MSS) or Base Station (BS)*: The connection of each cell with other cells and fixed hosts is established through a wireless interface that called mobile support station or base station. In addition this device can communicate with one MU through its wireless interface. Each BS has the address of whole cells, and when one MU of one cell enters into another cell, BS in first cell should modify the address of this MU so that the processing results of corresponding transactions of this MU send to its new address [1, 3].
- *Database Server (DBS)*: This part is necessary to incorporate full database functionality [5]. Each DBS can be reached by any BS or FH directly. But MUs communicates with a DBS only through BSs [6].

III. QUERY PROCESSING IN MOBILE DATABASE SYSTEMS

A. Types of Queries

Definitely, transaction processing is based on mobile database architecture in our proposed method. Generally there

are two types of queries in mobile environments. *Non-Location Related Queries (NLRQ)* and *Location Dependant Queries (LDQ)*. In fact, a NLRQ is a traditional query whose answer does not depend on locations and none of the predicates and attributes used in it are location related [2]. (e. g. Retrieve the names of hotel employees). Contrary to NLRQs, LDQs have at least one location related predicate or attribute [6].

B. Location Dependant Query

There are two types of location dependant query:

- *Location Aware queries (LAQ)*: which has an explicit indication of location (e.g. Select the names of hotels in Isfahan).
- *Location Dependant Queries (LDQ)*: The location value in these queries is not explicitly known when the query is asked. Their answers are dependent to MU's location; i.e. the mobility affects their processing (e.g. Retrieve the nearest ATM). In order to provide the answer to the query, first we have to know the location of the query issuer. When we find out the issuer's location (Location Binding), the query becomes location-aware [2].

C. The Architecture of The LDQ Processing

In the Mobile Network Services, the wireless network provides the wireless interface and the operator has the responsibility of supplying the location of the MU to authorized parties with the help of the Location Service (LS). We put the LS box in traditional mobile computing architecture and assume the location of the client is either provided by the network or by the device (GPS). MUs can communicate with LDISM, send query and view the results with the help of the interface [7].

Figure 2 shows architecture of the Location Dependant Queries processing.

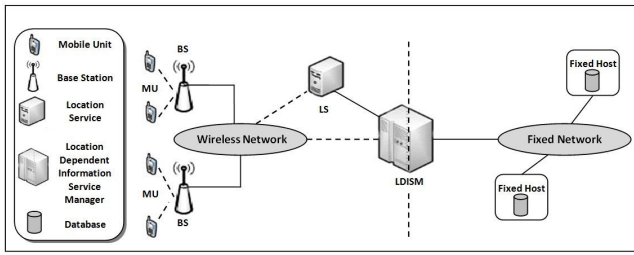


Figure 2. The architecture of the LDQ processing

IV. PROPOSED METHOD

A. Multipurpose Smart SIM Card

By designing a new chip that by helping of mobile database integrates the information of all the smart electronic cards and also authentication information of each person in itself we can solve many of the problems of smart electronic cards and their resultant troubles. This chip called multipurpose smart SIM card.

Multipurpose smart SIM card has been embedded in mobile phone and it is available all the time. We can use this SIM card by mobile database architecture as a national certification card, driving license, passport, people's medical history, fuel card, subway card, electronic election and some other applications.

B. Implementation and Integration

The practical implementation of this plan is required to provide appropriate telecommunications infrastructure and policies to integrate smart cards into the multipurpose smart SIM card. That should be established and provided by the authorities in charge. However, the proposed smart SIM card should be have major elements to integrate smart cards information. These elements are shown in figure 3 that are:

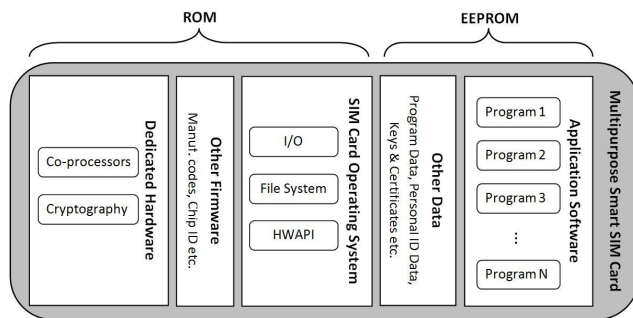


Figure 3. Multipurpose smart SIM card architecture

- **Central Processing Unit (CPU):** Is core of the smart SIM card that performs information processing.
- **Read Only Memory (ROM):** For carrying operating system and built-in programs loaded at the time the SIM card chip is manufactured. The memory of such a SIM card should be have a greater capacity in comparison with conventional SIM card, because

should be stored greater volume of data to integrate smart cards information.

- **Electrically Erasable Programming Read Only Memory (EEPROM):** Is non volatile memory that normally used for storing data and SIM card applications.
- **Random Access Memory (RAM):** Is writable and volatile memory and hence is only used for temporary storage.
- **SIM Card Operating System:** For controlling execution of application software, loading of new application program files, reading and writing of data to memory, and general low-level functions, such as power on and power off. In other words, the SIM card operating system is used for accessing the CPU, EEPROM, RAM and ROM.
- **File System:** This part may be part of the SIM card operating system that manages how data is stored and how programs on the SIM card can read and write to the EEPROM. To integrate information of smart cards into the multipurpose smart SIM card, this file system should be designed and implemented appropriate to provides information exchange.
- **Application Software:** This application runs on the SIM card CPU SIM card applications run in much the same way as regular personal computer.
- **Dedicated Hardware:** To deliver certain functions more securely and efficiently (e. g. cryptography, co-processors etc.).

C. File System Organization

As mentioned above, to integrate information of smart cards into the multipurpose smart SIM card, this SIM card should be have an appropriate and efficient file system. As shown in Figure 4, the file system of a multipurpose smart SIM card is organized in a hierarchical tree structure, composed of the following three types of elements:

- **Master File (MF):** The root of the file system that contains dedicated and elementary files.
- **Dedicated File (DF):** A subordinate directory to the master file that contains dedicated and elementary files.
- **Elementary File (EF):** Is a file that contains various types of formatted data, structures as either a sequence of data bytes, a sequence of fixed size records, or a fixed set of fixed size records used cyclically.

The Global System for Mobile Communication (GSM) standards defines several important dedicated files immediately under the MF: DF_{GSM}, DF_{DCS1800}, and DF_{TELECOM}. For the MF and these DFs, several EFs are defined, including many that are mandatory. The EFs under DF_{GSM} and DF_{DCS1800} contain mainly network related information respectively for GSM 900 MHz and DCS (Digital Cellular System) 1800 MHz band operation. EFs for U.S. 850 MHz and 1900 MHz bands are

found respectively under those DFs as well. The EFs under DF_{TELECOM} contain service related information [11].

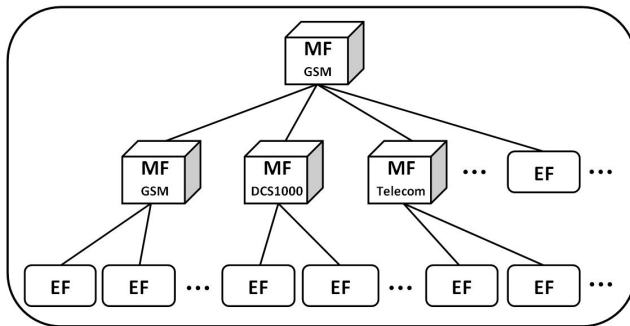


Figure 4. Multipurpose smart SIM card file system

After implementing this plan, people's communications and transmission of required information with banks, governmental centers, shops, gas stations, subways, and airports can be done using this smart SIM card. So, people will be able to perform several tasks like bank transactions, electronic shopping, paying the cost of vehicle fuel, cost of insurance, paying tolls of highways, paying the cost of public transportation systems in addition to use communicative facilities of their own mobile phones using this SIM card. Also if the data like certification information, driving license, passport and medical records have been saved in a database then all the governmental centers such as police offices, airports and hospitals can do their inquiries through this SIM card and using mobile database.

D. Authentication of Users

The first step to provide security is to recognize and authenticate the user's identity that wants to use database information. In this part we introduce two methods to authenticate in this system that the second method is our proposed method.

- Current method for user authentication in mobile phones is based on the use of 4-8 digit Personal Identification Numbers (PINs) [12]. In other words, authentication can be performed by a password which each user enters through his/her mobile phone. The implementation of this method is simple but if the mobile phone is theft or illegal people find its password it may have many hazards. Also this method may be threatened by destructive and pervasive software and it can be passed in fact we can prevent relatively by using a mechanism like figurative security codes that nowadays it is considered for most websites (e.g. yahoo) in order to user entrance.
- In second method in order to authenticate of users, we capture a photo from user's face and send it to fixed host, then we perform authentication according to received picture and the pattern which has been recorded in database from that user.

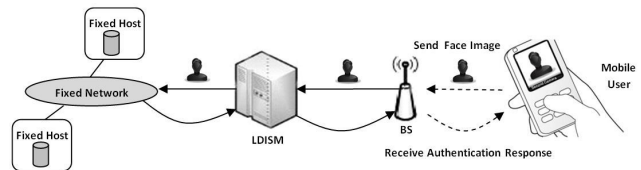


Figure 5. Authentication of users with image processing

Second method has high security and can solve the problems of the first method. But it needs an image processing system and the mobile phone of user should be equipped with second camera so that it can capture the image of user's face.

In addition to these methods, using new technologies and techniques such as PKI (Public Key Infrastructure) SIM or Biometric Techniques, It is possible to provide higher levels of trust and security for authentication. PKI SIM is an enhanced SIM card, which incorporates a digital certificate. This certificate is used to authenticate the user, so no username/password credentials are needed. Furthermore, it can be utilized as a digital signature and document signing in M-Police service. Biometric authentication techniques including fingerprint recognition, retinal scanning, hand geometry scanning, handwriting and voice recognition can be used too. These techniques are all based on the physical properties of a person [12].

E. Advantages and Disadvantages of Proposed Method

The advantages of this smart SIM card include: The ability of finding the nearest and the least occupied ATM or card reader device, high availability, easily transportation and facility in using, low waiting in the queues of ATMs or card reader devices, not need to carry so many smart cards, the ability of using as people's medical records, insurance booklet, driving license, passport, national card, fuel card, subway card and etc.

We can mention some possible disadvantages of this method: Low network bandwidth, power limitation (low power of batteries), and frequent disconnection, less reliability, higher probability of failure (loss, damage), latency and asymmetry in the communications (client-server vs. server-client) losing or thievery of SIM card, burning of SIM card, requiring an individual SIM card for each person and lack of necessary telecommunications infrastructure. [1, 4, 12].

V. CASE STUDY

In order to illustrate a practical application of the multipurpose smart SIM card, we designed a problem named finding the nearest and least occupied ATM. This case study will be described in three steps.

First step: Assume that a user is moving in the street and wants to reach to the nearest and least occupied ATM. He/she can perform a query by using his/her mobile unit and receives the address and distant of the nearest ATM according to his/her current location and considered distance radius from network. By using this query, even user can be aware of the number of waiting people in queue in order to use that ATM.

$$\prod_{atm-address, distance, count} \sigma_{(closest(here, radius))} (ATM)$$

User can receive the information of the nearest ATM according to his/her current location at given distance radius using a query as follow, (e.g. 3 items) and then select the most appropriate item according to the location of devices, their distances and the number of waiting people in queue.

$$\prod_{atm-address, distance, count} \sigma_{(TOP3closest(here, radius))} (ATM)$$

The result of above query is sent to user as follows:

TABLE I. RESULT OF QUERY

ATMs	Address	Distance	Count
ATM 1	Street No. 12	5km	4
ATM 2	Street No. 16	12km	6
ATM 3	Street No. 5	1km	2

This query is a location dependent query that its response depends on the user's location and his/her movement after sending query; it means that the responses are changing by user's movement continuously. Fixed host first should ask the current location of query's source from BS in order to respond to this query, so that it can send the result to the location of that mobile unit. If during performing the transactions of that query, user moves from one cell to another cell, the BS in first cell should modify the address of this mobile unit so that the result of query can be send to its new address.

Second step: After receiving the obtained results of query by user, he/she can send a turn request in order to use given ATM according to his/her current location in proportion to the nearest ATM and the number of waiting people. Host (ATM) allocate the requested turn to user and sends it to user through wireless network so that he/she would have its own turn and then approach to given ATM and use it. Also host should save the password of a person who has requested turn accompanied with turn number in its database that it can perform authentication while using ATM by that user and makes sure that the turn number of a person who is using ATM really belongs to his/her and he/she has not used another person's turn. Authentication performs by accommodating a password which user enters when he/she uses ATM and a password which the host has saved while requesting turn in its database.

Third step: When user reaches his/her given ATM, he/she should wait until his/her turn number is read by ATM, so he/she can use ATM. ATM calls turn numbers sequentially and wait for each turn one interval period (e.g. 15 seconds). If given person does not refer to ATM in that period, it will call the next turn number. If the user reaches given ATM late and his/her turn has been passed, he/she should send a new request turn for that ATM device again. In fact in introduced approach, just for getting money, there is a need for ATMs and doing above steps. Other bank operations like money transmission, bill payment, shopping services and etc. can be performed without requirement of ATM and through mobile unit and wireless connection with network.

CONCLUSION AND FURTHER WORKS

One of the requirements of electronic governments is to provide and distribute smart cards for different applications which cause each person should carry several and this issue can makes problems that we have mentioned in this paper. Our proposed method integrates the information of all the smart electronic cards into a multipurpose smart SIM card. This SIM card works based on mobile database architecture. Some of the advantages of this method include not need to carry various smart electronic cards, not waiting in the queues of several card readers, ease of use, high availability and easily carrying. This SIM card also can be used as a driving license, passport, medical history, electronic election, fuel card, subway card and some other cases.

One of the main challenges of this plan is to provide appropriate telecommunications infrastructure in order to practical implementation and developed the multipurpose smart SIM card system. Of course, the authors in future work will create a lab environment to simulate and test multipurpose smart SIM card system. Other critical issues of this system is warranty the reliability of wireless connections, security mechanisms in order to prevent the penetration of hackers and illegal users that we can study these challenges in the future. Also the detailed studying of other aspects of using this method (e.g. using as medical applications, electronic election, driving license, fuel card, etc.) can be a subject to future researches.

REFERENCES

- [1] D. Barbara, "Mobile computing and databases - a survey", IEEE Transaction on Knowledge and Data Engineering, vol. 11, no. 1, pp. 108-117, 1999.
- [2] A. Y. Seydim, M. H. Dunham, and V. Kumar, "Location dependent query processing", 2nd ACM International Workshop on Data Engineering for Wireless and Mobile Access, pp. 47-53, 2001.
- [3] M. H. Dunham and A. Helal, "Mobile Computing and Databases: Anything New?", SIGMOD Record, vol. 24, no. 4, pp. 5-9, 1995.
- [4] J. Li, Y. Li, M. T. Thai, and J. Li, "Data Caching and Query Processing in MANETs", Pervasive Computing and Communications, vol. 1, no. 3, pp. 169-178, 2005.
- [5] V. Kumar, "Mobile Database Systems", NJ: John Wiley & Sons INC., 2006.
- [6] M. Tarafdar and M. S. Haghjoo, "Location Privacy in Processing Location Dependent Queries in Mobile Database Systems", 5th IEEE International Symposium on Telecommunications, pp. 181-186, 2010.
- [7] W. Xinhua and L. Li, "Location Dependent Continuous Queries Processing Model Based on Mobile Agent", 9th IEEE International Symposium on Distributed Computing and Applications to Business, Engineering and Science, pp. 224-227, 2010.
- [8] S. Ilarri, E. Mena, and A. Illarramendi, "Location-Dependent Query Processing: Where we are and where we are heading", ACM Computing Surveys, vol. 42, no. 3, pp. 1-73, 2010.
- [9] UNCTAD, "Information Economy Report", United Nations, New York and Geneva, 2005.
- [10] Department of Finance and Deregulation, "National Smart Card Framework", Australian Government Information Management Office, 2008.
- [11] 3GPP, 2005a, "Specification of the Subscriber Identity Module - Mobile Equipment (SIM - ME) interface", 3rd Generation Partnership Project, TS 11.11 V8.13.0 (Release 1999), Technical Specification, 2005.
- [12] H. R. Firoozy-Najafabadi and S. Pashazadeh, "Mobile Police in Mobile Government", 5th IEEE International Conference on Application of Information and Communication Technologies, pp. 118-122, 2011.

Estimation models of competition and complementarity within communication technologies

Nurilla Mahamatov^{a,b}, Suk Won Cha^b

^aDepartment of mathematical and natural science, Turin Polytechnic University in Tashkent, Niyazova 17, Tashkent, Uzbekistan

^bDepartment of Mechanical and Aerospace Engineering, Seoul National University, Gwanakro599, Gwanakgu, Seoul, Republic of Korea 151744
Tel: +82-010-6874-5665, Fax: +82-2-889-6205
E-mail: mahamatov@yahoo.com

Abstract: The question of complementarity and competition often comes up when analyzing and comparing communication technologies. A corollary is that if the technologies are complementary they will be able to co-exist peacefully - otherwise they will engage in a battle of technological dominance. The communication services in this analysis include: Internet, mobile cellular phone and fixed line telephone service in Commonwealth of Independent States (CIS). Modified Diffusion Models used for analyzing of complementarity and competition dynamics within technologies. Analyses of estimation results and factors associated with ICT diffusion dynamism can play significant role to provide policy recommendations for these countries and for other developing countries to achieve the desired pervasiveness of the ICTs.

I. INTRODUCTION.

Early applications of diffusion models addressed primarily durable goods market such as TV set, refrigerator, etc. These models were single-product models concerned only with the sales growth for a single product [1]. The often-criticized insufficiency of them is that new products or technologies are not adopted into a completely new market without any other existing similar or related technology. The existence of other products and technologies may affect, no matter positively or negatively, the sales of a new product. Although those first-purchase models succeeded in illuminating the curve of diffusion at first, due to the increasing diffusion potential of technologies resulted from the increasing population, existence of other products or upgraded function of technology, constant potential has become the conceptual limit [2]. However, connection quality and number portability discrepancies between the two services are fading and substitutability may increase over time due to continued price declines and greatly improved connection quality. Furthermore, upgraded functions and improvements of mobile services will outpace those of fixed service.

The idea for creating a model of ICT diffusion has emerged from the need to identify the factor or group of factors that has been predominant in a case of a particular country. The Open model for ICT diffusion was built on the data from the CIS countries; Russia, Tajikistan, Uzbekistan, Azerbaijan, Kazakhstan and Kirgiz [3]. It has some similarities with the models for ICT diffusion in China, India, and Latin America. The differences from the other models are its openness, the

number of the principal factors, the non-rigid classification of the factors in categories, and their interaction. The data so far confirms many of the assumptions incorporated in the models. In the model, one of the objectives is to emphasize the cross-fertilization between the three different categories of factors and the feedback, which is not necessarily recursive.

The relationship between competition and complementarity in diffusion is particularly emphasized in the literature on the diffusion of innovations: in many circumstances increasing the number of technologies leads to faster diffusion of innovations [4].

The academic literature containing econometric analysis of the demand for mobile communications is limited. Most of the contributions rely on data aggregated to the metropolitan or country level. Hausman estimates the elasticity of aggregate subscription to cellular service in the 30 largest US markets over the period 1988–1993 [5], Ahn and Lee estimate demand for mobile access in Korea using more recent wireless subscription data for 64 countries [6].

Only recently has research appeared that examines fixed-mobile complementarity. Ahn and Lee find that the number of Korean mobile subscribers is positively correlated with the number of fixed-line disconnects, but negatively related to the number of new fixed-line connections, suggesting net complementarity between the two services. This pattern occurs even while the stock of fixed lines is positively correlated with the number of mobile subscribers, offering evidence that the two services are complements.

II. EXPERIMENTAL.

For the purposes of this paper, we define a technology to be technically complementary to another, if it can be used to provide a service that cannot be provided with the other technology. On the contrary, we define a technology to be technically substitutive to another, if it can be used to provide a service that can also be provided with the other technology.

To be useful in the analysis, we first have to define the concept of technical complementarity. In basic economics, as opposite to substitutes, two goods are considered to be complements if the demand for one decreases as the price of the other increases. In mathematics and e.g. set theory, the complement of a set includes all elements that are not in the set. For analyzing technical complementarity, the latter

association is perhaps more appropriate

Figure 2.1. Illustrates three technologies (A, B, and C), which are mapped into a graph based on their performance and capabilities in two different dimensions. Often, dimensions such as mobility and accessibility have been used when visualizing technologies capabilities. Figure 2.1a) shows a situation where the three technologies truly complement each other, i.e. are completely non-overlapping. In this case, the technologies each have their own specific application areas, in which they face no competition from the other technologies. Figure 2.1b) on the other hand, illustrates a situation in which the three technologies each have their own technical strengths and weaknesses compared to the other two, but are also partly overlapping (i.e. substitutive) with each other.

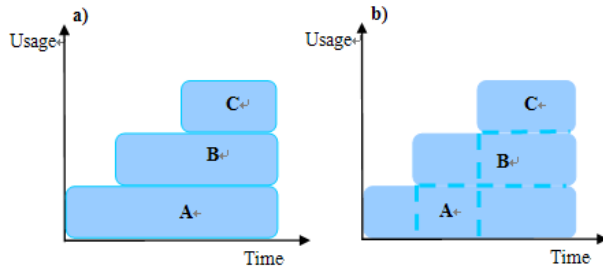


Figure 2.1: Illustration of full and partial technology complements

For the estimating model parameters several estimators are used. They are ordinary least squares (OLS), non-linear least squares (NLLS), three stage last squares (3SLS), full information maximum likelihood (FIML) and seemingly unrelated regressions (SUR). First, the OLS estimator was used for determining the values of the parameters of these models. Then, using these OLS-generated parameter values as initial values, the same model parameters were again estimated by using NLLS estimator. These models were specified implicitly by listing the dependent variable and independent variables after the name of the estimation method (OLS, 3SLS, FIML, SUR). Although this method of specifying a model is convenient, in order to estimate nonlinear models, we have to be more specific about the form of our equations. Two procedures in TSP estimate general nonlinear models with additive disturbances: LSQ and FIML. LSQ is a minimum distance estimator that can be used to compute nonlinear single equation least squares, nonlinear two-stage least squares, nonlinear multivariate regression, SUR (seemingly unrelated regressions), nonlinear three-stage least squares. FIML obtains full information maximum likelihood estimates for a nonlinear simultaneous equation model whose disturbances are jointly normally distributed. Both LSQ and FIML can be used on linear as well as nonlinear models. All statistical analyses in this study were done using the statistical software called Time Series Processor, version 4.5 (TSP 4.5). The t-statistics reported by TSP were used for assessing the significance of the parameters, while the goodness-of-fit (R^2 or adjusted R^2) or root mean square errors (RMSE) reported by TSP were used for selecting suitable models.

Extended Bass Model of diffusion

In light of the foregoing, an empirical analysis of the mechanism governing diffusion, complementarity and competition Dynamics inside ICT market is attempted by utilizing the diffusion models identical to respective diffusion dynamics [7]. From previous chapter we can conclude, inside ICT the empirical evidence of complementarity and competition between technologies and services, for analyzing this needed modified model. The estimation of Bass model parameters (p , q and m) may be obtained by using empirical data and the equation (1) in a classical regression analysis procedure. This procedure leads to a good fit even if the diffusion process is not yet completed [8]. In Extended Bass model we take account of the diffusion of the ICT market reflecting the interaction among telephone, mobile phone, and the Internet. In the formula, r_{ij} means the degree of the influence of service j on service i (telephone - 1; mobile phone - 2; Internet - 3).

Extended Bass model can be modified:

$$\frac{dy_i}{dt} = (p_i + q_i \frac{y_i}{m_i} + \sum_{j \neq i}^3 r_{ji} \frac{y_j}{m_j})(m_i - y_i) \quad (1)$$

Where:

$$\frac{dy_i}{dt} - \text{rate of adopters at time } t$$

y_i - cumulative number of adopters, p_i - coefficient of innovation, q_i - coefficient of imitation, m_i - market potential.

Subscript $i=1, 2, 3$ denotes an individual country, subscript j and $k=1, 2, 3$ specifies each telecommunications service. In addition, r_{jk} measures the interaction between two services. If the coefficient gets a positive sign, service j and service i am termed complements. A negative sign, on the other hand, shows that they are considered substitutes.

A good fit may be obtained when diffusion has reached the middle of the whole process, which is around the inflexion point of the curve.

For different product diffusion processes, p and q are comparable only if the frequency (e.g. monthly, annually) of observed data is roughly the same. On the other hand, this condition does not hold for the parameter m , as it can be compared even if data are collected in different frequencies.

Extended Logistic Model of diffusion

In Extended logistic model analysis, we take account of the diffusion of the ICT market reflecting the interaction among telephone, mobile phone, and the Internet. In the formula, r_{jk} means the degree of the influence of service j on service k (telephone: 1; mobile phone: 2; Internet: 3). Let $Y_{ij}(t)$ denote the penetration rate in each service market j . In addition, b is defined as the dependent variable affected by explanatory variables such as the other technologies. Let m_{ij} denote the market potential share over total population,

$$Y_{ij}(t) = \begin{cases} 0 & t < \tau_{ij} \\ \frac{\sum m_{ij}}{1 + \exp\left\{-a_{ij} - (b_{ij} + \sum_{k \neq j} r_{ik} Y_{ik}(t))(t - \tau_{ij})\right\}} & t \geq \tau_{ij} \end{cases} \quad (2)$$

subscript $j=1, 2, 3$ specifies each telecommunications service

and is the entry time of each service. In addition, measures the interaction between two services. If the coefficient gets a positive sign, service j and service k are termed complements.

A negative sign, on the other hand, shows that they are considered substitutes. Equation (2) should not be treated separately because some explanatory variables are jointly determined with the dependent variable. In this perspective, we make use of estimate simultaneous equation by nonlinear FIML (Full Information Maximum Likelihood) estimator, LSQ estimator, SUR estimator and 3SLS estimator.

LSQ is used to obtain least squares or minimum distance estimates of one or more linear or nonlinear equations. SUR obtains seemingly unrelated regression estimates of a set of nonlinear equations. 3SLS obtains three stage least squares estimates of a set of nonlinear equations. Which estimator can give a more reasonable results will be analyzed.

To make the estimation, the TSP (Time Series Processor) 4.5 program, one of computer programs that are widely used for the computations, is executed.

Extended Gompertz Model of diffusion

The advantage of this process is that we are able to estimate Gompertz adoption curves for different ICT services and countries that characterize the reality of income related digital division. The advantage of this method over aggregated approaches is that the planner can observe and estimate consumer activities that conform well to our existing understanding of the underlying economic processes of demand; it is therefore easily described to non-technical policy makers. A second advantage to this method is that the choice of segmentation variable is left open to the analyst to decide.

There is another frequently used S-shaped diffusion function, the Gompertz diffusion curve. It is described by the function:

$$Y_{ij}(t) = m_{ij} * \exp(-\exp(-a_{ij} - (b_{ij} + \sum_{k \neq j} r_{ijk} Y_{ik}(t))(t - \tau_{ij})) \quad (3)$$

Where; $Y_{ij}(t)$ - is penetration rate in each service, m_{ij} - is market potential, $i = 1, 2, 3$ - is individual country, $j = 1, 2, 3$ - is specifies each service, τ_{ij} - is entry time of service, r_{ijk} - measure the interaction between two services.

The Gompertz curve is inherently non-linear and therefore requires estimation using non-linear estimations techniques, that is, to use non-linear least squares (NLLS) as this method should provide least bias in the parameter estimates. The Gompertz models were estimated in the TSP (Time Series Processor) 4.5 program, using its own NLLS estimator. Some attention is also required in the assessment of the statistics that are provided.

III. RESULTS AND DISCUSSION

Given several different types of approximation models, can be used to select the one with the best accuracy value. Table 3.1 shows the average error in estimating the $RMSE$ (Root mean square error) and R^2 for the three Diffusion models functions. The $RMSE$ of an estimator is the expected value of the square of the "error". The error is the amount by which the estimator differs from the quantity to be and estimated by

equation (4).

We can see that the estimation error of the $RMSE$ is different for all three models, also shows the average error in estimating R^2 . For these "Best Model" selection problems, the estimation error for the $RMSE$ in "Model II" is quite low then in other two models. In other hand estimation of the R^2 are high in this "Model II" then in other two models.

We can conclude that "Best Model" selection is more reliable with estimated results of $RMSE$, R^2 and select "Model II" (Extended Logistic Model) as the "Best Model".

For the robustness of the results and good fitness of observations and estimations, several estimators are used. They are: SUR (seemingly unrelated regressions), 3SLS (three-stage least squares) and FIML (full information maximum likelihood). Previous method also used for "best estimator" selection and presented in Table 3.2.

SUR estimator is more reliable by estimation results of $RMSE$ and R^2 . In this study Extended Logistic Model with SUR estimator will be used for the following parameter estimations. $RMSE$ can be estimated by equation:

$$RMSE = \sqrt{\frac{1}{T} \sum_{t=1}^T (Y_t - \hat{Y}_t)^2} \quad (4)$$

Table 3.1: The Model parameters analysis by $RMSE$ and R^2

Parameters	MODEL I (Extended Bass)	MODEL II (Extended Logistic)	MODEL III (Extended Gompertz)
Fixed Telephone	38086	17855	59417
Mobile	6987	8301	11826
Internet	24328	21666	13618
R^2	0.69	0.96	0.74
Total ($RMSE$)	69402	47823	84861

Table 3.2: The estimators analysis by $RMSE$ and R^2

Parameters	3SLS	FIML	SUR
Fixed Telephone	31058	23876	14131
Mobile	20245	36152	31731
Internet	68273	19192	25352
R^2	0.78	0.82	0.94
Total ($RMSE$)	119576	85380	65054

Extended Diffusion Models include new parameters r_{ij} which measures the interaction among services and estimation for all services is simultaneous. In additional r_{12} - is measures the interaction fixed to mobile, r_{13} - fixed to internet, r_{21} - mobile to fixed, r_{23} - mobile to internet, and r_{31} - internet to fixed, r_{32} - internet to mobile.

All parameters estimated by using Extended Bass Model, Extended Logistic Model and Extended Gompertz Model, using of different methods gives possibility for best Model selection. Table 3.3 Extended Logistic Model parameters of the diffusion in CIS. All three Modified Diffusion Model estimates are similar and shows relationship among communication services in these countries mostly is complementary. According to each country and service, parameter does not have definite sign, which indicates that

countries have no distinctly uniform trend. However, it appears likely that fixed line service affected Internet service positively among four countries (Azerbaijan, Kyrgyzstan, Russia and Uzbekistan) in the same manner, except Kazakhstan and Tajikistan where affected negatively. Conversely, Internet service has tendency to affect fixed line service negatively in case of Azerbaijan and Uzbekistan. It means that the xDSL service based on fixed line becomes popular in these countries as time passes. Coefficients such as r_{12} , r_{31} and r_{32} are statistically insignificant in some countries. On the other hand, r_{13} and r_{23} are statistically significant positive value at level of 5%, which means that growth of fixed phone and mobile telephone leads to more Internet usage in the same manner. However, more fixed phone subscribers result in a lower level of mobile phone penetration level because of the negative value of r_{12} in case of Russia and Tajikistan. In addition, mobile phone and Internet have the relationship of substitution with each other.

Interaction among communication services in CIS countries which are mostly complements (Figure 3.1). The framework is considered to be widely applicable for analysis of future telecom technologies.

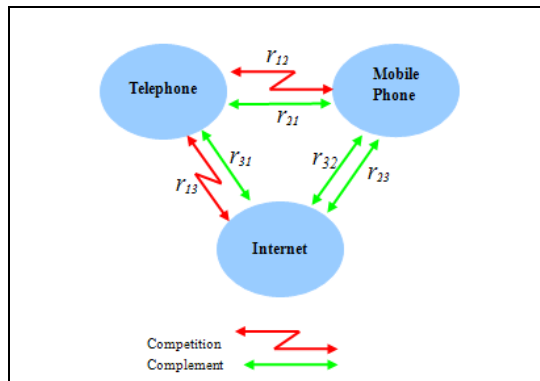


Figure 3.1 Interaction among technologies in CIS

The results of Extended Logistic Model are highly significant than other models result, (Higher number of R^2 and less number of $RMSE$). Considering all these comparisons, the Extended Logistic diffusion model can be selected for the rest of the analyses of the diffusion of the Telecommunication services in this study.

The number of main telephone lines in CIS has shown a minimal growth rate for several years, in Tajikistan registered a negative growth rate in fixed lines and only four countries had growth rates of fixed line. Despite these developments the region remains in terms of fixed line penetration. An important issue, and a potential threat to fixed line operators, is the fixed to mobile substitution, that is, the use of mobile phones instead of fixed phones for calls or access. Fixed to mobile substitution may take different forms, and is not always clearly measurable, and interpretations and predictions about its scale and impact vary. Fixed to mobile substitution seems to be an even greater issue in the lower-income regions of CIS and fixed line growth rates remain low even in countries where penetration levels have not reached saturation. The Commonwealth of Independent States (CIS), which at a

modest 3.8 percent in 2003, showed the highest fixed line growth rate, had only 21 fixed lines per 100 inhabitants. Much higher growth rates in the mobile market, 89 percent in CIS in 2003, suggest that many users choose to have access to a mobile phone only.

In 2003 most of the CIS countries - Kyrgyzstan, Tajikistan, and Uzbekistan - had penetration levels fewer than five percent. The CIS average remains at a very low 17.1 percent. At the same time the region with the lowest mobile subscribers per 100 populations has by far the highest growth rates, an average of 99.4 percent. Tajikistan leads at 441 percent growth rate between 2001 and 2003. Another example for continued market opportunities is Russia; market with the most potential in terms of population more than doubled the number of mobile subscribers in 2004, from 36.5 million to 74.4 million. The country's poor fixed line network and lack of infrastructure investment has opened up opportunities for three large operators. High growth in 2004 has allowed Russia to overtake Germany, France, Spain and the UK and to become the largest mobile market in Europe and CIS. However, the country's overall mobile penetration rate of close to 52 percent conceals a major national digital divide, with the majority of subscribers concentrated in urban centers. Internet penetration levels show that the digital divide that separated CIS from the rest of OECD is greater than in mobile and fixed lines.

Table 3.3: Extended Logistic Model parameters of the diffusion in CIS (Equation 2)

Parametr	RUS		TAJ		UZB	
	Est.	StdErr	Est.	StdErr	Est.	StdErr
a_1	0891	19202	6.0431**	19294	2.0793**	0814
b_1	.1093	.1238	-.8920*	.3937	.0728	.0397
r_{11}	-.101E-08	.313E-08	.133E-04	.799E-05	-.953E-07	.849E-06
r_{31}	.8206E-08	.360E-07	.102E-03*	.490E-04	.108E-05	.478E-06
R^2	0.99		0.69		0.96	
a_2	-9.5505**	.6468	-34.2161**	6.4552	-3.9149**	.7358
b_2	.8484**	.2583	7.0681**	2.0083	-3.0680	1.5986
r_{12}	-.169E-08	.546E-08	-.12E-04**	.450E-05	.194E-05	.929E-06
r_{32}	.3085**	.650E-09	-.17E-03**	.624E-04	.267E-06	.108E-06
R^2	0.99		0.99		0.99	
a_3	-6.0233**	.2770	-10.4321**	1.0842	-5.1366**	.4933
b_3	.8625**	.1469	5.4441**	.7223	-2.0234*	.8320
r_{13}	-.97E-08**	.339E-08	-.17E-04**	.260E-05	.14E-05**	.468E-06
r_{23}	.14E-08**	.167E-09	.26E-05**	.687E-06	-.50E-09*	.215E-07
R^2	0.99		0.99		0.99	

Parametr	AZB		KAZ		KYR	
	Est.	StdErr	Est.	StdErr	Est.	StdErr
a_1	3481**	.0430	1.5579**	1.790	1.9082**	2310
b_1	.0121**	.390E-02	-.13312**	.0322	-.999E-02	.06853
r_{11}	.15E-06**	.124E-07	.120E-06**	.267E-07	.8254E-04	.747E-04
r_{31}	.794E-07	.539E-07	.111E+08**	.19E+08	-.133E-04	.128E-04
R^2	0.98		0.97		0.77	
a_2	-8.4243**	1.5706	-12.476**	2.2604	-9.2523	9045
b_2	.4939	.6048	.8875	.6680	-.213E-02	.5646
r_{12}	.4656	.535E-06	.217E-06	.1450	.174E-05	.114E-05
r_{32}	-.30E-06*	.100E-06	-.11E-05**	.105E-06	.524E-06*	.202E-06*
R^2	0.96		0.99		0.99	
a_3	-11.176**	.4260	-6.7429**	1.1649	-5.8783	3.4912
b_3	-2.7407**	.1183	1.9574**	.4467	-1.0011	3.2654
r_{13}	.478615**	.199E-06	-.65E-06**	.154E-06	.432E-05	.735E-05
r_{23}	-.65E-06*	.281E-07	.2169E-06	.091E-06	.554E-08	.390E-06

Notes: Statistically significant ***at level of 1%, **at level of 5%, *at level of 10%

We have investigated the factors which determine the diffusion of the Telecommunication services in CIS countries by estimating the Extended Logistic model of technology diffusion with data on fixed line, mobile cellular and Internet for the years 1993–2005.

The results show an uncertain pattern with regard to the effects of each country's interaction between communications services. This study explains the diffusion of communications services by considering the interaction among technologies.

Accurate technology diffusion forecasting is extremely important for today's managers and technology planners. In addition to these theoretical contributions, the results of this study have policy implications for governments to develop interventions at appropriate times to induce faster diffusion of Telecommunication services. Our results should aid domestic telecommunications companies in planning export strategies. For example, in the CIS mobile phone service and the Internet are complementary services that simultaneously enhance the growth prospects of each other, companies should export mobile phones together with products related to the Internet.

These countries have an initial condition that differentiates them from developed countries. They still suffer from poor fixed line telecommunications infrastructure, due to lack of investment during the centrally planned period. Therefore, these countries have an open choice regarding the type of the technology that could be utilized in the future, at least for narrowband telecommunications. Mobile services seem like a superior choice for this purpose.

The majority of studies on substitutability or complementarity among telecommunications services have paid attention to the interaction of fixed phone service and mobile phone service, but the present study assigns fixed telephone and mobile phone service as well as Internet service to the category of telecommunications services by incorporating the effects of the Internet in the telecommunications service market. Given the estimated results we may conclude that initial hypotheses regarding the diffusion are showed to hold in most specifications. Thus, income per capita, access costs, infrastructure development and existence of legislation related to ICT explain the diffusion process best; all these factors appeared to be highly statistically significant.

The present analysis should help demand forecasters to plan the provision of plant and build capacity when a new telecommunications service appears in the ICT market, and the analysis has significance for the market entry strategy applicable to a network industry distinguished by greater investment in initial infrastructure

Finally, this research is expected to be useful in devising a plan to strengthen telecommunications market competitiveness and in constructing policy applied to market environment change.

- [1] Bass, F. M. (1969) A new product growth model for consumer durables, *Management Science*, 15, 215–227.
- [2] J.W. Woodlock, L.A. Fourt, Early prediction of market success for grocery products, *J. Mark.* 25 (1960).
- [3] International Telecommunications Union (ITU) (2006) *World Telecommunications Indicators 2006*. Geneva, ITU.
- [4] Chen, C. and Watanabe, C. (2006) Diffusion, substitution, and competition dynamics inside the ICT market: The case of Japan, *Technological Forecasting and Social Change*, 73(6), 731-759.
- [5] Hausman, J. (1999). Cellular telephone, new products and the CPI. *Journal of Business and Economic Statistics*, 17(2), 188–194.
- [6] Ahn, H., & Lee, M.-H. (1999). An econometric analysis of the demand for access to mobile telephone networks. *Information Economics and Policy*, 11, 297–305.
- [7] Islam, T., Fiebig, D. G. and Meade, N. (2002) Modeling multinational telecommunications demand with limited data, *International Journal of Forecasting*, 18, 605-624.
- [8] Mahajan, V., Muller, E. and Bass, F. M. (1990) New product diffusion models in marketing: A review and directions for research, *Journal of Marketing*, 54, 1–26.

Power-Aware Topology Generation for Application Specific NoC Design

Suleyman Tosun
Ankara University
Faculty of Engineering
Computer Engineering Dept.
Ankara, Turkey.
Email: stosun@ankara.edu.tr

Yilmaz Ar
Ankara University
Faculty of Engineering
Computer Engineering Dept.
Ankara, Turkey.
Email: ar@ankara.edu.tr

Suat Ozdemir
Gazi University
Faculty of Engineering
Computer Engineering Dept.
Ankara, Turkey.
Email: suatozdemir@gazi.edu.tr

Abstract—Network-on-Chip (NoC) is a new alternative approach to bus-based and point-to-point communication methods to design large System on Chip (SoC) architectures. Designing a power-aware irregular topology for a NoC based application is a challenging problem due to its high complexity. This paper tackles at this problem and presents a genetic algorithm based topology generation algorithm (GATGA) for NoC architectures aiming to minimize the power consumed by the communication among tasks of the application. Our experiments on multimedia benchmarks and randomly generated graphs show that the proposed algorithm achieves considerable improvements over the existing topology generation algorithms in terms of communication overhead and power consumption.

I. INTRODUCTION

System on Chip (SoC) design complexity has increased significantly in recent years. More and more cores have been added on systems with each technology generation. This technology improvement allowed the designers building the entire system on a single chip by placing several processing cores, memory elements, and I/O units. Bus-based communication is not feasible for SoCs due to several reasons, such as concurrency, clock skew, and long bus lengths [1], [2]. The traditional network concept opened a new on-chip communication infrastructure called Network-on-Chip (NoC) [3], [4]. NoC is a good alternative to bus-based communication structure for systems with many numbers of components since it increases the bandwidth as a result of parallel execution of the components. Furthermore, NoC solves the clock skew problem by using routers as decoupling elements in the system.

NoC architectures can be constructed by using either regular topologies or irregular (custom) topologies. There have been several regular topologies tested for NoC architectures [5] such as mesh, torus, double torus, butterfly, fat trees etc. The regular topologies are easy to construct and they are reusable because of their unchanged architecture nature. However, the applications cannot be optimized well on regular topologies. The irregular topologies are designed to be application specific. As a result, they give more optimization opportunities than their regular counterparts for objectives such as power consumption, area, number of routers in the system.

Even though the advantages of irregular topologies are clear enough, there is still need for scalable topology generation algorithms. This study is a step towards fulfilling these needs. In this study, we propose a genetic algorithm based topology generation algorithm (GATGA) for Network-on-Chip architectures. The main contributions of this work are threefold:

- Compared to existing solutions, we present a new, fast and easy to implement genetic algorithm for the topology generation problem. In our algorithm, we apply evolutionary functions on both network architecture and node-to-router connections. In this way, we can have topology generation and mapping at the same time.
- Our work applies different synthesis order when designing the NoC system as opposed to earlier methods. It first generates the topology of the network and then places the components on the chip area. That is, floor planning comes after the generation of the architecture.
- We provide extensive experimental results to compare GATGA to the existing algorithms. The results show that GATGA outperforms the existing algorithms by achieving up to 32.30% decrease in total communication and up to 17% power reduction on real multimedia benchmarks.

We organized the rest of the paper as follows. Section II presents the related work. While Section III gives the problem definition, Section IV explains the proposed algorithm. Section V demonstrates the experimental results. Finally, concluding remarks are presented in Section VI.

II. RELATED WORK

Most of early work focused on designing NoC architectures on regular topologies such as mapping algorithms [6], [7], [8] and NoC design tools [9], [10]. These studies have opened the NoC architecture design area by showing the advantages of NoC over its classical alternatives. There have been work that focused on irregular topologies [11], [12], [13] as well. In their pioneering work, Srinivasan et al. [11] use two-step integer linear programming (ILP) based irregular topology construction method. In the first step, their ILP based method determines a suitable floor planning for the given application. Then, in the next step, it finds the connections between router and nodes. Their method obtains very promising results. However,

as they state in their work, ILP based methods takes very long execution times. This disadvantage makes ILP based irregular topology generation inapplicable for application containing huge number of nodes. In [12], Chang and Chen present a two-step topology generation heuristic, called PATC. In the first step, they cluster the nodes based on the communication weights. That is they try to place highly communicating nodes to the same cluster. After cluster optimization in second step, they construct the topology based on the clusters. In [13], authors use a better clustering method than PATC and achieve better results. In this work, we compare GATGA with all the above mentioned topology generation methods. Although tremendous research has been conducted on the area of NoCs, there are still needs for algorithms for the application specific NoC topology construction. This paper is a step towards fulfilling these needs.

III. PROBLEM DEFINITION

Let $G(V, E)$ represent the core flow graph (CFG) where V is the set of all vertices and E is the set of edges. Each v_i in V represents a node and each $e_{i,j}$ represents an edge between the nodes¹ v_i and v_j . The weight $w_{i,j}$ of the edge $e_{i,j}$ shows the communication cost between nodes i and j . Fig. 1 depicts an example CFG. The set of routers of the topology is represented by R and the set R consists of r routers. In our router set, each router R_i has the same number of ports p where $1 \leq i \leq r$. Given the above specifications, the topology generation problem can be defined as follows:

Given a $G(V, E)$ and a set of routers R , determine a topology that satisfies:

- 1) Each router R_i is connected to at least one other router R_j , if $r > 1$,
- 2) Each v_i is connected to a router,
- 3) There exists a communication path between every v_i and v_j where $e_{i,j} \in E$, and
- 4) The total communication cost (or the total power consumption) is minimized.

When we calculate the communication cost, we determine the transferred data amount among routers. In our work, we aim to minimize the total communication cost in an attempt to minimize the power consumption of the system. This is because the communication cost is proportional to power consumption as discussed in [7]. One may opt to this assumption since the communication cost and power consumption of the system does not scale linearly. There are different criteria affecting the power consumption besides the power consumed by the network. Since our aim is minimizing the power consumed by the network resources, we focus on minimizing the number of bits travelling over the network resources. We show the improvements in both communication cost and power consumption in Section V and we observed that the improvement percentages are not the same for two parameters. However, improving the communication cost greatly improves the power consumption values.

¹In this paper, we use the terms core and node interchangeably.

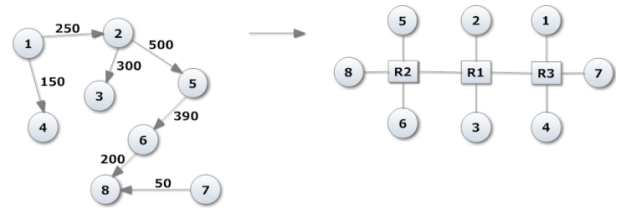


Fig. 1. An example CFG (left) and the generated topology (right).

Fig. 1 shows the topology generated from the CFG using GATGA. As can be seen from this figure, each router is connected to at least another router, each node is connected to a router, and there is a path between each communicating node pair.

IV. THE PROPOSED ALGORITHM: GATGA

The proposed Genetic Algorithm based Topology Generation Algorithm (GATGA) is based on evolutionary computing. Each possible topology is represented as an individual string where router ports are the genes of the individual. GATGA is composed of 3 major steps, namely; *initialization*, *crossover*, and *mutation*. In the initialization step, GATGA randomly generates P_{size} valid individual strings and uses P_{low} individuals that have the lowest communication cost. In the crossover and mutation steps, GATGA randomly selects P_{rand} individual pairs and performs multiple runs of crossover and mutation operations over the selected P_{rand} individual pairs to minimize the communication. P_{size} , P_{low} , P_{rand} , the number of cores (n), and the number of ports (p) on each router are the system parameters that are determined using the input application. In what follows, we explain the details of GATGA.

A. Determining the Number of Routers

First task of our algorithm is determining the minimum number of routers required by the system. After finding the number of routers utilized in our system, we can create individuals that represent valid topologies. We have to determine the number of routers first since the string size of each individual is equal to the number of available router ports. Having the number of nodes (n) from our input CFG and the number of ports of the routers (p) in our router library, we can calculate the minimum number of routers (r) using the following formula, $n \leq pr - 2(r - 1)$. We derive this formula by assuming that we use exactly $r - 1$ links to connect r routers. That is if we have r routers, each of which having p ports, there can be pr available ports. Since each router must be connected to another router via at least one link, $(r - 1)$ communication links are needed to connect all routers. Since each communication link consumes 2 ports, then to connect all routers, $2(r - 1)$ ports are required. The remaining ports can be used to connect cores to routers.

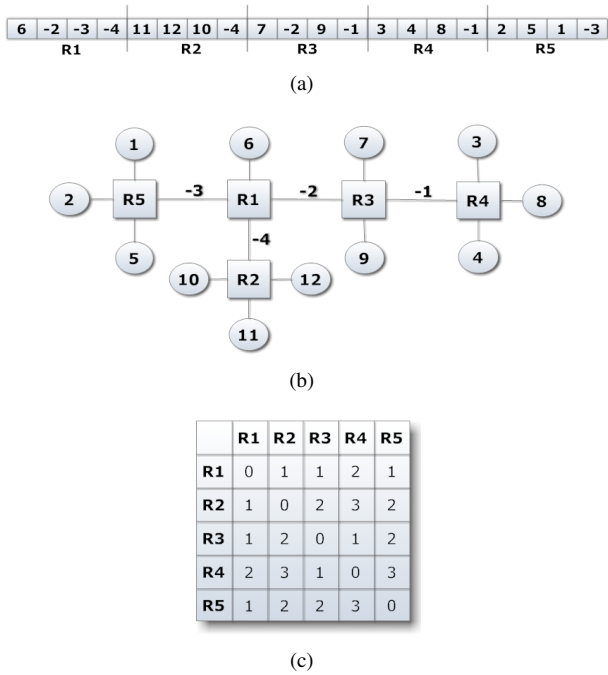


Fig. 2. An example chromosome (a), its corresponding topology (b), and its distance matrix (c).

B. Illustrative Example

Fig. 2(a) illustrates an example individual string, representing the topology given in Fig. 2(b). The positive numbers in genes represent the node number that is connected to this port. For example, in Fig. 2(a), node 11 is connected to port 1 in router R2. It should be noted that the individual string must contain each positive number exactly once because each node should be placed in exactly one position in the topology. The negative numbers in genes indicate the communication links between routers. For example, in Fig. 2(a), -2 represents the link number 2 that connects routers R1 and R3 in specified ports. Also note that the individual string must contain each negative number twice, representing two ports the links connected to. If there is no link or node assigned to a port, we place the number 0 to the corresponding gene. If the number of 0s in a string is more than one, we can add an extra link to the topology for each pair of 0s. In this work, we only use the minimum number of links.

In our example designs, we use the optimum number of routers, resulting in at most one empty port of the network. However, we can easily extend this work by adding links to the empty ports, which results in alternative communication paths between communicating nodes. In the example in Fig. 2, there are 12 nodes and 4 communication links connecting 5 routers. Therefore, all 20 ports are used by nodes (12 nodes) and links ($2 \times 4 = 8$ ports for 4 links).

C. Initialization

In the first step of GATGA, individual strings are created to form an initial population by randomly assigning nodes

and communication links to genes. We select the population size based on the given graph size. In our experiments, we selected our population size as $P_{size} = 1000$. For bigger sized graphs, we can increase this limit to increase the probability of obtaining better results with a cost of CPU execution time.

After we randomly generate an individual string, we check if this string represents a valid topology. To have a valid topology, each node must be assigned exactly once in the string and each link number must be assigned to exactly two genes, each of which belong to different routers' portions. We then add the generated valid individual string to the population list L_{ind} . The first step of GATGA continues until the size of L_{ind} reaches to P_{size} .

During the population generation, we have to calculate the fitness function (i.e., the cost) of each individual. We use the cost of each individual as selection criteria among several candidates. Our cost is based on the total communication cost of the topology that the individual represents. To calculate this cost, we create a distance matrix based on the minimum hop distance between any two routers. Finding the minimum hop distance between two routers is the well known shortest path problem and we use Dijkstra's algorithm to solve it. Fig. 2(c) shows the distance matrix of the individual presented in Fig. 2(a). Using the distance matrix, we compute the total communication cost among the routers ($Cost_i$) as the fitness value for each individual i . We then sort L_{ind} based on $Cost_i$ values of individuals and select the first P_{lowest} individuals having the lowest cost among remaining candidates. In our experiments, we selected P_{lowest} as 200. After creating the initial population, GATGA generates a second generation population using genetic operators: crossover and mutation.

D. Crossover

In the crossover operation, we copy L_{ind} to a new list L'_{ind} and we randomly select P_{rand} pairs of individuals from L_{ind} . We apply crossover operation for each selected pair by randomly swapping two routers in each pair of individuals. In most cases, this crossover operation produces two invalid individuals. Thus, we must repair the faulty assignments that make these individuals invalid. There can be four cases that make an individual string invalid. We list these cases and repair procedure of GATGA in the following paragraphs.

1) *Multiple nodes in the string*: If the string representing a solution contains one node more than once in the gene list, we say that this individual is invalid. In our crossover operation, we swap only one router from each string. In this case, a node may reside at most twice in a string. GATGA randomly selects one of these nodes and assigns a "0" to it. By doing this, it opens a slot (i.e. an empty port) to the node that is not assigned because of multiply assigned ones.

2) *Missing node in the string*: If a node of the CTG is not included in the string, that string represents an invalid topology. In such a case, we randomly assign the missing node/nodes to the genes that contain "0". Since we occupy these places in the previous step, the number of 0's in the string is always more than the number of missing nodes.

3) *Excess links*: Each link of the topology is represented by the same negative number in the string. The two negative numbers must be in different routers, meaning that the link connects these two routers. After a swap operation, at least one link value is added to the string, which may result in having three same negative numbers in the string. In our definition of irregular topologies, a link is allowed to connect only two routers. In such a case, the string representing the solution becomes invalid. GATGA eliminates one of three negative numbers by replacing its value with "0".

4) *Broken links*: When we swap two routers, one from each individual, one of the negative values may be moved to the other pair. In this case, this link value presents three times in an individual, which we mentioned in Case 3, while the same link value is contained only once in the other individual. In this case, both individuals are invalid. While we replace the exceeding numbers with "0", we randomly assign the missing values to the genes that have the value "0". By doing this, we connect the broken link to another router. In this step, we place an extra control to GATGA not to place the same negative numbers to the same router ports more than once.

The crossover operation is finalized by creating the distance matrix and computing $Cost_i$ for each individual i in L'_{ind} .

E. Mutation

After the crossover operation, GATGA performs mutation operation over L'_{ind} by selecting P_{rand} individuals randomly. In each selected individual, it swaps two genes randomly. There are six possible mutation outcomes as follows:

- 1) An empty port (i.e. "0" values) can be swapped with another empty port.
- 2) A node port can be swapped with an empty port.
- 3) A link port can be swapped with an empty port.
- 4) A node port can be swapped with another node port.
- 5) A node port can be swapped with a link port.
- 6) A link port can be swapped with another link port.

In the above procedures, the reproduced individual can be invalid only in cases 3, 5, and 6. In these cases, a link value is replaced to another router that may cause to have two same negative numbers in the same router. We prevent the occurrence of the case 6 by using the same control mechanism of GATGA we mentioned in crossover operation. However, for the cases 3 and 5, the network can be disconnected. We can check whether the network is connected or not, after we create the distance matrix. If there is no path between two routers, we assign the infinity value to the distance value of these routers in the distance matrix. Obviously, the cost of the disconnected networks will be infinity in such a case. Thus, we eliminate them from our population.

The mutation operation seems to be very simple and in fact it is very easy to implement. However, it is very powerful to create different types of architecture. Its attractiveness comes from the fact that it does not limit itself to only node swapping, which is general case for most of the evolutionary computing based topology generation algorithms. By selecting both nodes and links at the same time, GATGA can change

the architecture of the network (i.e. router connections) and node assignments at the same time. This enables us to evaluate different architectures with different node mappings.

After the mutations, GATGA creates the distance matrices for each individual and computes their costs. At the end of the mutation process, the list L'_{ind} is added to the original list L_{ind} resulting in a list of size $2P_{lowest} = 400$. L_{ind} is sorted again using $Cost_i$ values and only the P_{lowest} individuals that have the lowest $Cost_i$ values are kept in L_{ind} for the next iteration of crossover and mutation loops. GATGA performs the crossover and mutation P_{size} times and outputs the individual with the lowest cost as our solution.

V. EXPERIMENTAL RESULTS

In this section, we evaluate our topology generation algorithm GATGA by comparing it with previous work through several experiments. For this evaluation, we used two sets of CFGs: multimedia benchmarks and randomly generated graphs. We used five different multimedia benchmarks, namely; VOPD and MWD from [12], 263 Dec., 263 Enc., and Mp3 Enc. from [11]. In our experiments, we assumed that we have routers with four ports, which is the mostly accepted router type. For power consumption parameters, we adopted the power model for 100-nm technology given in [12]. In this model, the power consumptions of the routers are estimated as 328 nW/Mb/s and 65.5 nW/Mb/s for input and output ports, respectively. In addition, the link power consumption is estimated as 79.6 nW/Mb/s/mm.

We compared GATGA with 3 different topology construction algorithms presented in [11], [12], and [13].

A. Evaluation on multimedia benchmarks

Tables I and II summarize the results of the experiments conducted on multimedia benchmarks. From column two to four, Table I gives the total communication costs obtained by GATGA, TopGen, and previous work (ILP [11] and PATC [12], respectively). In column four of Table I, the first two rows are the communication cost results obtained by PATC and the remaining results are communication cost results generated by ILP. In columns five and six, we give the communication cost improvements of GATGA over TopGen and ILP-PATC, respectively. As can be observed from this improvement percentages, GATGA outperforms all three algorithms for all multimedia benchmarks, achieving up to 32.30% improvement over the best algorithm results among three algorithms. This due to the fact that GATGA does not limit itself to predefined clustered nodes. Thus, it improves the probability of obtaining different architectures. By doing so, GATGA can generate several node mapping configurations. However, TopGen and PATC first determine a cluster and bound themselves to one possible grouping of nodes. ILP first generates the floor planning of the cores. To our observations, this limits to place heavily communicating nodes to the same router. In our work, we did not consider the floor planning in the first phase. The floor planning phase comes after the topology is generated.

TABLE I
COMMUNICATION COST IMPROVEMENTS OF GATGA AGAINST PREVIOUS METHODS.

Application	Communication Cost (Mbits/sec)			Improvement of (1) over (2) (%)	Improvement of (1) over (3) (%)
	GATGA (1)	TopGen (2)	ILP/PATC (3)		
VOPD	1289	1872	1875	31.14	31.25
MWD	672	800	1024	16.00	34.38
263 Dec	1.482	1.492	1.492	0.67	0.67
263 Enc	62.888	87.715	87.715	28.30	28.30
MP3 Enc	2.630	3.885	4.778	32.30	44.96

TABLE II
POWER CONSUMPTION IMPROVEMENTS OF GATGA AGAINST PREVIOUS METHODS.

Application	Power Consumption (μ Watt)			Improvement of (1) over (2) (%)	Improvement of (1) over (3) (%)
	GATGA (1)	TopGen (2)	ILP/PATC (3)		
VOPD	2491.44	2999.32	3001.90	16.93	17.00
MWD	1026.09	1137.60	1332.72	9.80	23.01
263 Dec	9.018	9.02	9.02	0.02	0.02
263 Enc	145.37	166.99	166.99	12.95	12.95
MP3 Enc	8.79	9.88	10.66	11.03	17.54

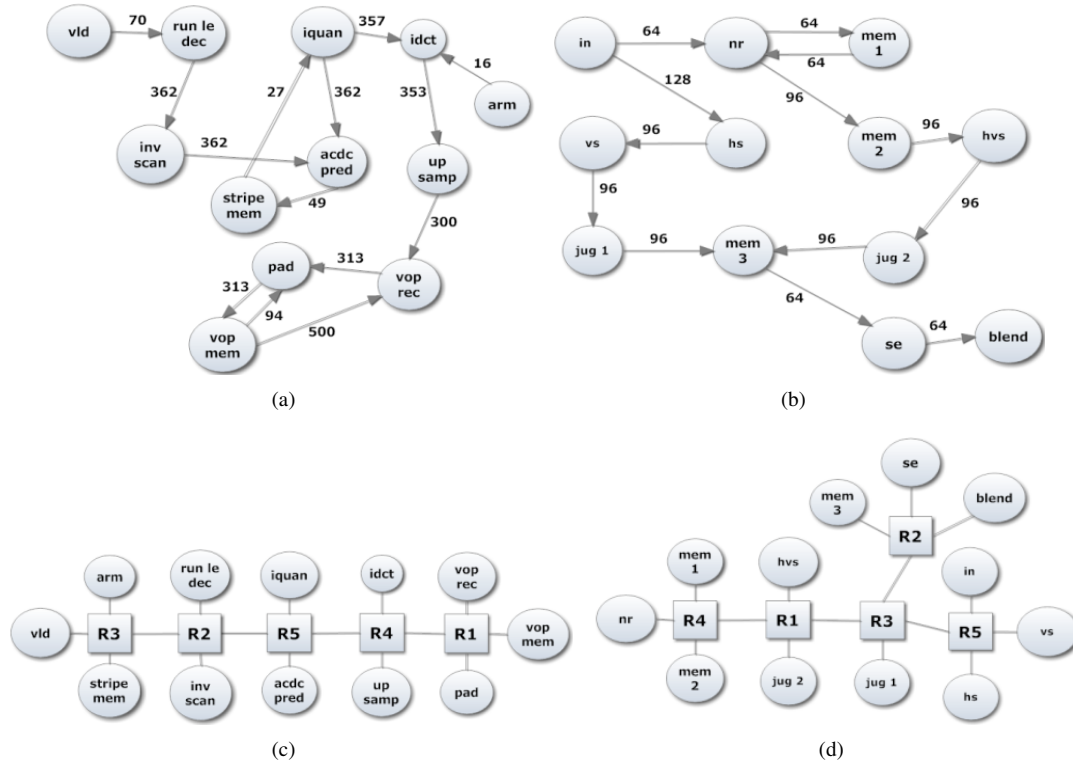


Fig. 3. CFG of VOPD (a) and MWD (b) and their generated topologies (c) and (d), respectively, by GATGA.

As we stated earlier, our ultimate goal is minimizing the power consumption of the created NoC system. To achieve this goal, we stick to the fact that, in communication, the power is consumed only when the bits from one core to the other are transferred. Therefore, minimizing the communication cost of the system reduces its power consumption. Of course, the power consumption improvement will not be the same as communication cost since the power consumed on the routers and the links are not the same. For example, if two nodes are placed on the same router, they consume some amount of

power although we take their communication cost as 0. If they are placed on different routers, we add both the power consumed on routers and on links. Similarly, their communication costs will increase. Using the 100-nm technology parameters, we calculate the power consumed by the NoC communication except computation units. We give these results for GATGA, TopGen, and ILP-PATC in columns 2, 3, and 4 of Table II, respectively. The last two columns of Table II show the power consumption improvements of GATGA over TopGen and ILP-PATC, respectively. GATGA outperforms three algorithms for

all benchmarks and achieves up to 17% power consumption improvement over the best cases of three algorithms. In fact, as we expected, we observe same success order of communication costs and power consumption values for four algorithms. Due to page limit concerns, we only present the generated topologies of VOPD and MWD benchmarks in Fig. 3 as a reference.

B. Evaluation on randomly generated graphs

The number of nodes in the above mentioned multimedia benchmarks ranges from 12 to 15. To evaluate the scalability of GATGA, we also evaluated it with randomly generated graphs, whose nodes ranges from 5 to 160. We compared GATGA and TopGen in these experiments since TopGen achieves better results than ILP and PATC as reported in [13]. We evaluated them based on communication cost improvements and execution times.

In Fig. 4(a), we show the power consumption improvement of GATGA over TopGen on randomly generated graphs. We observe the highest improvement increase on the graphs between 10 to 15 nodes. For each node increase, the improvement increase lessens. Fig. 4(b) shows the execution times of GATGA and TopGen algorithms. As we can see from this comparison, the execution times of GATGA suddenly bounce to very high values when the number of nodes increases from 80 to 160. Additionally, if we check the improvement from 80 nodes to 160 nodes in Fig. 4(a) it is around 2%. From these two graphics, we conclude that GATGA cannot be used on the graphs with huge number of nodes when there is a time limit. However, it can still be used for the graphs with large number of nodes when time limits are not concern.

VI. CONCLUSION

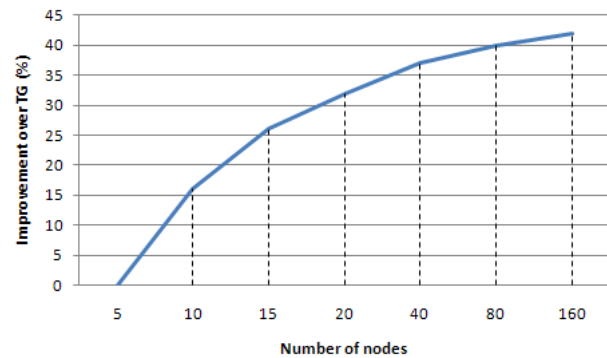
In this study, we present a genetic algorithm based topology generation algorithm (GATGA) for Network-on-Chip architectures to reduce the power consumption of the system. The results show that our algorithm makes a significant improvement compared to earlier studies. Our approach reduces the power consumption up to 17% for real multimedia benchmarks. Experiments conducted on graphs with varying number of nodes show that GATGA achieves better topologies minimizing the communication cost of the system with an extra cost of execution time.

ACKNOWLEDGMENT

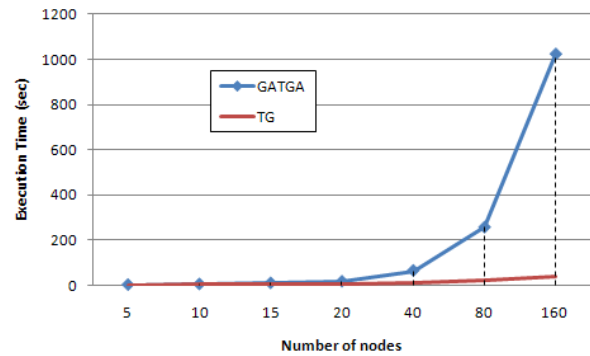
This work is supported by Scientific and Technological Research Council of Turkey (TUBITAK) under grant number 108E233 and EU COST Action IC0805 ComplexHPC.

REFERENCES

- [1] J. Henkel, Wayne Wolf, and Srimat Chakradhar: "On-chip networks: A scalable, communication-centric embedded system design paradigm," Proc. Of. 17th International Conference on VLSI Design, pp. 845-851 (2004).
- [2] H. G. Lee, N. Chang, Ü. Y. Ogras, and R. Marculescu: "On-chip communication architecture exploration: A quantitative evaluation of point-to-point, bus, and network-on-chip approaches," ACM Trans. Design Autom. Electr. Syst. 12(3) (2007).



(a)



(b)

Fig. 4. Power consumption improvements (a) and execution times (b) of GATGA over TopGen.

- [3] W. J. Dally and B. Towles, "Route Packets, Not Wires: On-Chip Interconnection Networks," Proc. Design Automation Conference, Las Vegas, Nevada, USA, 2001.
- [4] L. Benini and G. De Micheli, "Networks on Chips: A New SoC Paradigm," IEEE Computer, vol. 35, no. 1, pp. 70-78, Jan. 2002.
- [5] P. P. Pande, C. Grecu, M. Jones, A. Ivanov, and R. Saleh, "Effect of traffic localization on energy dissipation in NoC-based interconnect," ISCAS, pp. 1774-1777, 2005.
- [6] K. Srinivasan and K. S. Chatha: A technique for low energy mapping and routing in network-on-chip architectures," ISLPED, pp. 387-392, 2005.
- [7] J. Hu and R. Marculescu, "Exploiting the Routing Flexibility for Energy/Performance Aware Mapping of Regular NoC Architectures," Proc. of DATE'03, pp. 688-693, 2003.
- [8] S. Murali and G. De Micheli, "Bandwidth-Constrained Mapping of Cores onto NoC Architectures," Proc. DATE'04, vol.2, pp. 896-304, Feb. 2004.
- [9] A. Jalabert, S. Murali, L. Benini, and G. De Micheli, "XpipesCompiler: A tool for instantiating application specific Networks on Chip", DATE'04, 2004.
- [10] Z. Lu, R. Thid, M. Millberg, E. Nilsson, and A. Jantsch, "NNSE: Nostrium network-on-chip simulation environment," DATE 2005 University Booth Tool Demonstration, Munich, Germany, March 2005.
- [11] K. Srinivasan, K. S. Chatha, and G. Konjevod, "Linear-programming-based techniques for synthesis of network-on-chip architectures", IEEE Trans. Very Large Scale Integr. Syst. 14, 4 pp. 407-420, 2006.
- [12] K.-C. Chang and T.-F. Chen, "Low-power algorithm for automatic topology generation for application-specific networks on chips", IET Comput. Digit. Tech., 2008, Vol. 2, No. 3, pp. 239-249.
- [13] Ar, Y.; Tosun, S.; Kaplan, H.; , "TopGen: A new algorithm for automatic topology generation for Network on Chip architectures to reduce power consumption," AICT 2009,2009.

Wireless Platform for Multi-Channel RTD Measurements

Amer Atta Yaseen
College of Electrical and Electronic Techniques
Foundation of Technical Education
Baghdad, Iraq
E-mail: amer.atta@t-online.de

Abstract— The trend in monitoring and control systems has been to reduce the communications wiring across a plant. Eliminating the wired connection between a sensors and a host monitoring or control unit enhances the performances and the manageability of the system. This paper describes the practical design of wireless measurement platform, which uses Bluetooth communication and provides up to twelve RTD (Resistance Temperature Detector) measurement channels. The system contains RTDs connected to a microcontroller which controls also the Bluetooth module transmission, and a computer connected to a Bluetooth adapter. In this way, the RTD sensors data are collected and processed by the microcontroller, transmitted via Bluetooth to a host computer which monitoring and recording the RTDs temperatures measurements. The proposed wireless platform can be configured to measure other applications.

Keywords- wireless platform; Bluetooth; RTD.

I. INTRODUCTION

With the advent of the new low-cost wireless technologies, the functionality of RTD temperature acquisition systems can be enhanced eliminating the burden of cabled connection between the system itself and the data processing unit. Especially in the industrial environment, where every supplementary plug implies design and financial efforts, a wireless link between a RTD sensors acquisition module (sensor with signal conditioning and digital conversion) and a data processing unit (PC, Laptop or PDA) brings flexibility and robustness to the entire design [1].

Bluetooth is an open standard for short-range digital radio. It is designed to operate in the unlicensed ISM band. It has been developed to set-up “Pico” networks. Bluetooth uses a master/slave-based MAC protocol and enables low power consumption and short-range wireless connection between various electronic devices. Bluetooth is intended to replace the cables connecting portable and/or fixed electronic devices. The technology also offers wireless access to LANs, PSTN, mobile phone networks and the Internet [2]. Integration of Bluetooth and mobile products has been discussed in a study by Kirby [3]. It is already possible for Bluetooth to extend its application scope to 3G wireless communication systems [2]. The possibility of an industrial use for Bluetooth in data collection has been discussed in a study by Anderson [5] A home automation system, using Bluetooth, has been experimentally built up

with Sriskanthan et al [4]. Bluetooth using data transmission from inside vacuum chamber measurements [6] is an actual example of industrial usage. Also, the authors tested the interferences between Bluetooth and other wireless technologies [7], susceptible to be present in a home or industrial environment.

The system adapted in this paper is a low-cost and an easy-to-collect RTD (Resistance Temperature Detector) sensors data by the PIC18F4550 microcontroller and transfer them via Bluetooth to a host computer. The architecture of the mu system developed is detailed in the next sections.

II. HARDWARE DESIGN

A. Microcontroller Support Circuitry

The Microcontroller Support Circuit diagram is shown in Fig. 1 which represents the wireless platform board diagram. The PIC18F4550 microcontroller [8] offer cost-efficient solutions for general purpose applications written in C that use a real-time operating system (RTOS). Originally the PIC18F4550 was used due to its 13 ch 10bit ADC, EUSART (Enhanced Universal Synchronous Asynchronous Receiver Transmitter) embedded modules and smaller size. The microcontroller is powered from a 5V supply which is obtained via a 7805-type 5V regulator with a 9V input from a battery. An LED followed by a resistor is connected to the output of the voltage regulator as an on/off indicator for the microcontroller. The output node from the regulator is also shared by the Vdd pin on the microcontroller which is connected to the Vss pin (ground) through a 0.1 μ F decoupling capacitor. The same node is also connected to the Vpp pin through a 10k Ω resistor in series with a LED (a simple diode is sufficient) along with a switch that connects to ground to form the reset circuit. Through this the microcontroller can be reset by bringing Vpp down to ground. A crystal, along with two 22pF capacitors leading to ground, is connected to the OSC1 and OSC2 pins of the microcontroller to create a 20 MHz oscillator. Several tests were performed on the microcontroller to ensure proper functionality without any problem.

B. Analog-to-Digital Converter module

Analog to digital conversion is performed by the converter embedded in the microcontroller. The resolution

is 10 bits, with 1023 counts for the maximum level. There is only one analog-to-digital converter on the microcontroller, namely, only one channel can be converted at a time. An analog multiplexer connects the selected input to the holding capacitor and to the ADC. The input range is set by high and low voltage references. The microcontroller's internal voltages, $V_{dd} - V_{ss}$ (5V – 0V), were used as reference. The ADC is of the successive approximation type. Practical tests have revealed a maximal sampling frequency of 100 Hz.

C. RTD Temperature Sensors

RTDs from many manufacturers can be chosen for the proposed wireless platform system and a standard Class B RTD is chosen with an α of 0.00385. The reading of RTD sensors is made through the 12 channels 10-bit ADC embedded in PIC18F4550. The twelve signal conditioning circuits are used for each RTD as shown in Fig. 1, Assuming that the RTD is close to the microcontroller and the lead resistance errors are negligible. A 5 V constant voltage source type will be used to sense the temperature, from the platinum RTD characteristics,

at 0°C the resistance is $R_t = 100 \Omega$

at 100°C the resistance is $R_t = 171.5 \Omega$

By using a source resistance of $R_s = 1 \text{ k}\Omega$, and a source voltage of $V_s = 5 \text{ V}$, the voltage across the RTD element at either end of the operating temperature will be:

At 0°C , $V_T = 454 \text{ mV}$

At 100°C , $V_T = 608 \text{ mV}$

Using an operational amplifier with a gain of 5, the voltage range seen by the A/D converter will be $5 \times 454 = 2270 \text{ mV}$ to $5 \times 608 = 3040 \text{ mV}$. In case of 10-bit A/D converter will be used with a full-scale of 5 V, 1 LSB = $5000/1024 = 4.88 \text{ mV}$ but the input voltage range is $3040 - 2270 = 770 \text{ mV}$, or $770 \text{ mV}/100^{\circ}\text{C} = 7.70 \text{ mV}/^{\circ}\text{C}$. Thus, our system will be accurate to about 1°C .

D. Bluetooth Module

The Bluetooth module is functionally broken into three major sections: the baseband controller, flash memory, and the radio. The Bluetooth radio defines the requirements of the Bluetooth device operating in the 2.4GHz license-free Industrial, Scientific, and Medical (ISM) band. The baseband controller includes firmware for the host controller interface (HCI), which handles the communication with an external host (e.g., microcontroller, computer). The host and the baseband controller can communicate with each other for the transmission and reception of data through a Universal Asynchronous Receiver Transmitter (UART). The Bluetooth module used for this project is RBT-001 from Robotech [9]. The Main features of RBT-001 are Class 2 operation (nominal range up to 30m), Low power consumption, UART Command/Data Port supports for up to 921.6k baud rate, Integrated chip antenna, Small size (29x29mm). The RBT-001 is a SPP (Serial Port Profile) module which practically emulates a data serial transmission

the data received on the UART is sent further also on the UART interface. This circuit works together with other Bluetooth modules that supports the same SPP profile or can connect through the UART serial interface to a processor or direct to the system, depending of the application. Through an external processor or host (personal computer) all the available profile applications could be set to the SPP profile, for example: Dial Up Networking, Fax, LAN Access.

E. USART Setup and Bluetooth Connection

The major challenge, was establishing a Bluetooth link with a computer. However, before the link could be established the USART port of the PIC had to be configured. This port was responsible for the asynchronous serial transmission between the PIC and the Bluetooth module Through the USART, ASCII commands can be sent to RBT-001 module which enables it to perform a number of actions. This includes scanning the area for other Bluetooth devices, connecting to a specific device, disconnecting, etc.

RBT-001 Bluetooth module was connected to the microcontroller using only three wires of the serial interface, RX, TX and GND. Hence, software handshaking is performed using start and stop bits. One wire goes from the transmitting end of the PIC to the receiving end of the RBT-001 and vice versa for the other wire.

The host computer (PC or portable) has been incorporated with a Bluetooth 2.0 adaptor connected to the USB interface. To manage the Bluetooth connection, it is necessary to know whether the adaptor is compatible with SPP profile of RBT-001. The steps that are made are the following: the software for the Bluetooth adaptor is installed; the adaptor is connected to the PC's USB and the searching for Bluetooth devices is started; when the RBT-001 is found its name will appear as "Serial Port Device"; right click on the device name that was found and choose the option pairing; the PIN code of the device is inserted '0000' and OK is pressed. The adaptor's green LED is turned on, to show that the steps have been made correct.

For the primary experimental receiving of the data, the Microsoft HyperTerminal is used and set on the COM port to have the following parameters: 8 data bits, 1 stop bit, no parity and 921.6kbps.

F. Power Management

The microcontroller and RTD signal conditioning circuits are powered from a 5V supply which is obtained via a 7805-type 5V regulator with a 9V battery input. The 3V supply required by the Bluetooth module is obtained via a BA03T / FP regulator with 3V output and is driven from the 9V battery. The voltage of the battery also apply to the channel (13) after adjusted by 10K Ω potentiometer in order to obtain (0-5V) level of PIC18F4550 ADC module for the purpose of monitoring the battery voltage.

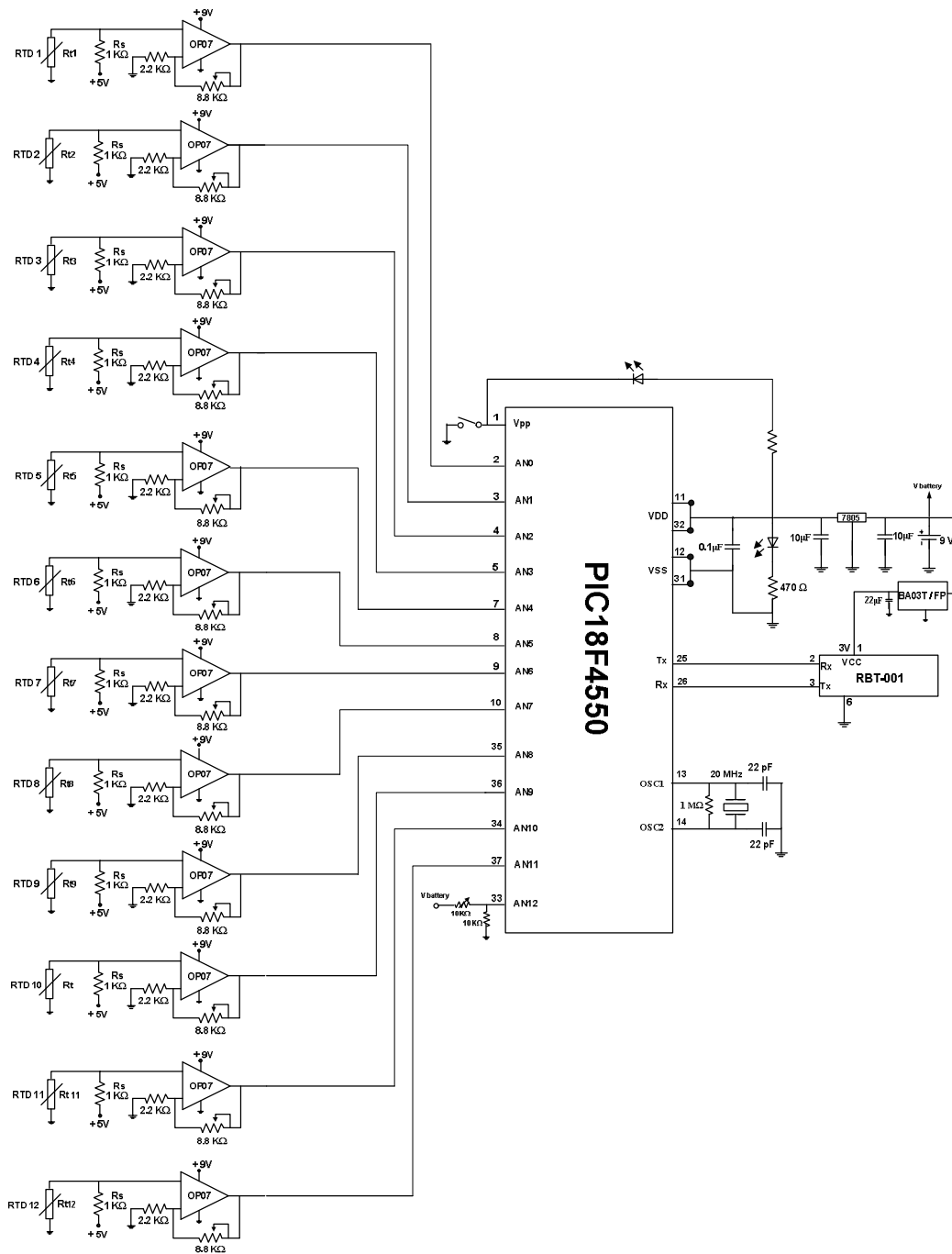


Figure 1. wireless platform board diagram

III. TEMPERATURE CALCULATION

A. RTD temperature resistance relationship

In practice, temperature-resistance relationship of the RTDs are approximated by an equation known as the *Callendar-Van Dusen* equation which gives very accurate results. This equation has the form:

$$R_t = R_o [1 + At + Bt^2 + C(t-100)^3] \text{----- (1)}$$

Where A, B, and C are constants which depend upon the material. Above 0°C, the constant C is equal to zero and we can re-write equation as:

$$R_t = R_o [1 + At + Bt^2] \text{----- (2)}$$

Thus, if we know the constants A and B and the resistance at 0C° then we can calculate the resistance at any other positive temperature using equation (2). However, in practice it is required to calculate the temperature from a knowledge of the RTD resistance. Equation (2) is a quadratic equation in t and it can be solved to give:

$$t = \frac{-RoA + \sqrt{Ro^2 A^2 - 4RoB(Ro - Rt)}}{2RoB} \text{ ----- (3)}$$

B. Temperature Calculation at Host Unit

The voltage across the RTD is converted into digital form and send via Bluetooth to the host unit. The host unit program calculates the resistance of the RTD (Rt) using equation (4) as follows ..

$$Rt = \frac{VT.RS}{Vs - VT} \text{ ----- (4)}$$

Where VT is the voltage across the RTD, Vs = 5 V, and Rs = 1 kΩ. Thus,

$$Rt = \frac{10^3 \cdot VT}{5 - VT} \Omega$$

For the used RTDs sensors

$$Ro = 100\Omega, A = 3.9083 \times 10^{-3} \text{ and } B = 5.775 \times 10^{-7}$$

The temperature is then calculated using equation (3).

$$t = \frac{-0.39083 + \sqrt{0.15274 - 2310 \times (Rt - 100) \times 10^{-7}}}{-1155 \times 10^{-7}} \text{ (5)}$$

If Rt is known, temperature can be calculated from equation (5).

IV. SOFTWARE DESIGN

A. The Firmware Programming

The program flowchart for the firmware of the microcontroller is shown in Fig. 2. The firmware program has been written in 'C' language, using the mikroC compiler developed by mikroElektronika [10].

The bluetooth_daq.c contains the functions, which initializes the microcontroller, acquires the sensors data and sends them to the host computer using the Bluetooth principle. When the application starts, first of all the microcontroller modules initiated as following:-

- USART is initiated using **Usart_Init** function.
- ADC is initiated using **Adc_Init** function.

These functions are called in the **Init_Modules** function. It sets the ADC module to the first channel selection, and empties the transmit buffer. On completing initialization, the ADC starts to digitizing the current selected channel. Wait for A/D conversion to complete, by

Polling for the A/D Conversion Status bit to be cleared and the high and low byte of the 10-bit conversion result are transformed into a line of characters which are stored in the **unsigned char buffer**. Selection of the next channel follows. The process will loop for the next twelve subsequent channels (in addition to battery voltage level monitoring channel (13)). After the total thirteen channels data packet have been written to the **buffer**, the 26 byte of data from **buffer** are sends byte by byte to RBT-001 via the USART in the background of a **for loop**, using the **Usart_Write** function and every packet starts with the character 'B' to indicate its beginning. The 26 byte data is arranged in ascending channel order with the arrangement of high byte followed by low byte for every channel as shown in Fig. 3, with this manner there is no need for appending channel identification since the location of the channel and byte is predetermined in the array. This way, the channel demultiplexing process at the PC side has been made simpler.

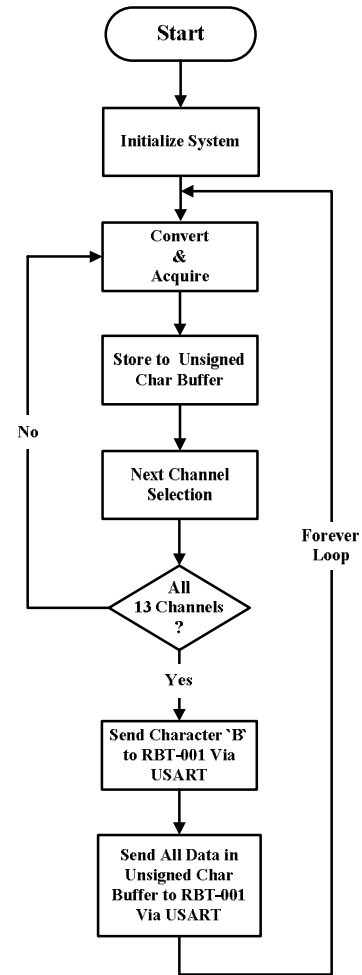


Figure 2. The firmware program flowchart.

character B	Channel 1		Channel 2		Channel 3		Channel...	Channel 13	
One Byte	High Byte	Low Byte	High Byte	Low Byte	High Byte	Low Byte	High Byte	Low Byte

Figure 3. Data packet of 27 bytes for transfer to PC via Bluetooth.

B. Host Computer Software

The host computer program is developed in Visual C++ environment on a Windows XP platform. The program replicates Microsoft HyperTerminal and handles the search, connection and disconnection of the link with the wireless platform board. Fig. 4 shows the host computer program Flowchart. Once the Bluetooth link between the wireless platform board and the host is established, the RBT-001 Bluetooth module appears as a virtual serial port on the host computer, the COM port was set to have the following parameters: 1 start bit, 8 data bits, no parity, 1 stop bit and 921.6kbps. The 921.6kbps is selected to match the maximum baud rate of the Bluetooth module. The Visual C++ program was successfully able to receive, decode data from the wireless platform board and immediately calculate the RTDs temperatures using equation (4 & 5), sorts and appends the appropriate data into 12 individual channel arrays and save it to 12 separate Excel files, as well as displaying it on the screen (in addition to the wireless platform battery voltage level monitoring channel (13)). Fig. 5 shows the host computer program user interface.

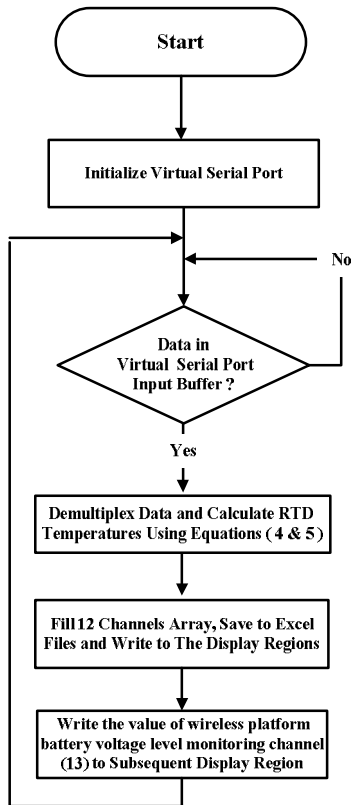


Figure 4. The host computer program flowchart.

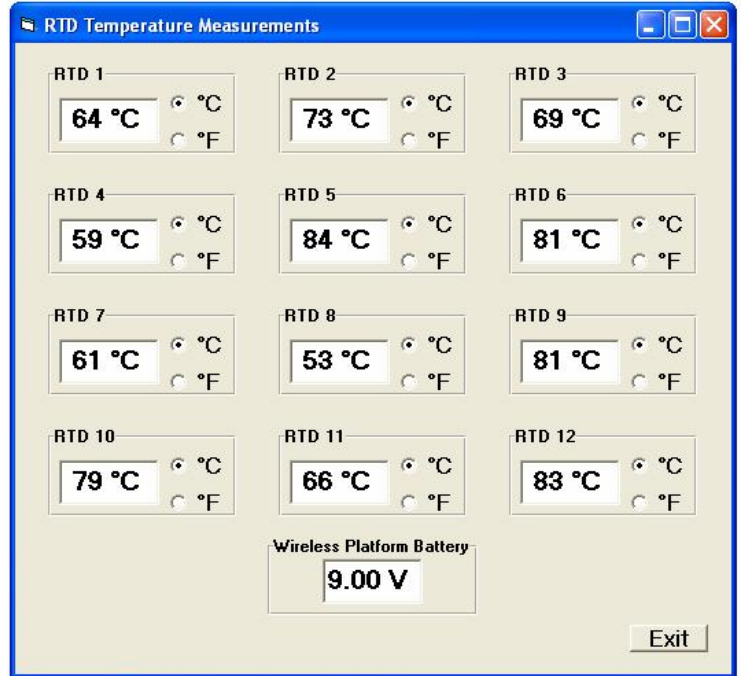


Figure 5. The host computer program User Interface.

V. RESULTS AND DISCUSSION

The communication between the acquisition board and the PC is succeeded using the Bluetooth module RBT-001, which practically emulates a serial communication. The data is acquired and sent to the PC has been incorporated with a Bluetooth 2.0 adaptor connected to the USB interface. The a 1 Hz sine wave with peak value of 5 V rectified by full-wave rectifier was used as the input signal with sampling frequency of 33Hz for the first test and 100Hz for the final test Fig. 6 and Fig. 7 shows a part of that acquired singles achieved by special visual C++ program was written and used for the test of this wireless platform system.

For each data point in Figures, a total of 2V bytes are transferred from the acquisition board to the host computer in one packet of data. This is because all thirteen channels of the remote acquisition card are actively transmitting their data, and the rectified sine wave is inputted to just one of the channels. When the two figures of the acquire rectified sine waves are compared, the changes are very significant and notable. Because of the higher sampling frequency (over three times greater than that of Fig. 6) Fig. 7 appears much more like a rectified sine wave. Both graphs; however, contain a significant amount of error, which is most notable near the signal peaks. The error is introduced into the signal at the microcontroller, during the analog-digital conversion. The increase in baud rate and sampling frequency seems to have large effect on this error.

The maximal sampling frequency is directly proportional to the baud rate at which operates the Bluetooth module (maximum 921.6kbps). Practical tests have revealed a maximal sampling frequency of 100 Hz. Further

improvements in the programming of the microcontroller will raise this limit. This operational mode is useful in systems where the data received from sensors is needed in real-time but consequent accuracy is not the main item. For future developments, data pre-processing (for example threshold detection) could be done directly in the microcontroller.

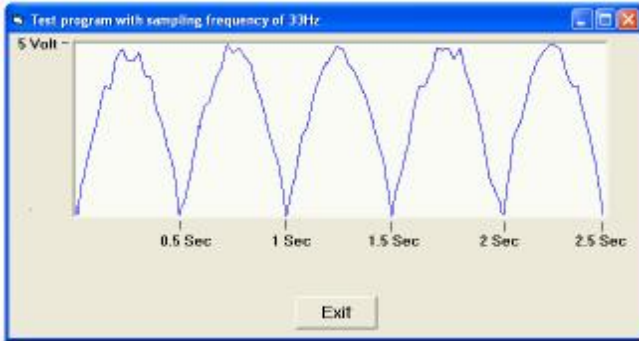


Figure 6. 1 Hz rectified sine wave sampled at 33 Hz.

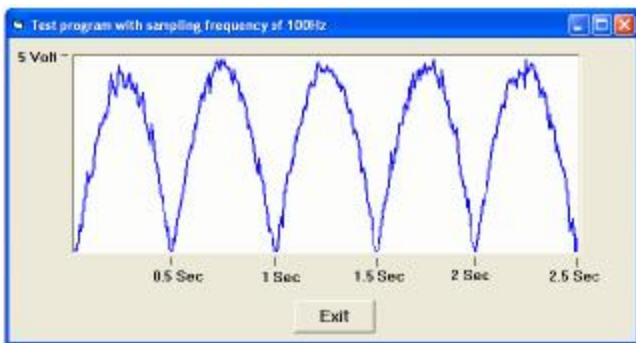


Figure 7. 1 Hz rectified sine wave sampled at 100 Hz.

VI. CONCLUSION

In this paper, a design of small-size, low-cost wireless platform for multi-channel RTD measurements equipment based on Bluetooth is introduced. The data of experiment indicates that this equipment has a good stability and precision. The realized user interface has a simple structure and it can be modified easily to display data acquired from numerous sensors. The demonstrative system can be extended for measuring different parameters of interest, like: humidity, gas concentration, presence, etc. Various indexes meet the designing request, and it can be used in the electrical measurement and testing area widely.

REFERENCES

- [1] Vlad Popescu, Iuliu Szekely "Wireless Data Acquisition System Using Bluetooth," Transilvania University of Brasov, 2005.
- [2] Marian Alexandru, " Remote Temperature Recording Using Bluetooth Technology," Acta Technica Napocensis Volume 48, Number 3, pp 27-30, 2007.
- [3] G Kirby, "Integrating Bluetooth technology into mobile products," Intel Technology Journal Q2, pp 1-8, 2000.
- [4] N. Sriskanthan, F. Tan, A. Karande, "Bluetooth based home automation system," Microprocessors and Microsystems 26, pp.281-289, 2002.
- [5] M Anderson, "Industrial use of Bluetooth," <http://www.connectblue.se>, 2001.
- [6] A. Murari, L. Lotto, "Wireless communication using detectors located inside vacuum chambers," Vacuum, pp.1-7, 2003.
- [7] M. Alexandru, M. Romanca, "Indoor Location Sensing and Tracking – Experiment-Supported Comparison of Bluetooth and Wireless Technologies," Proceedings of the 9th International Conference on Optimization of Electrical and Electronic Equipments, Brasov, Romania, vol. 4, pp.83-88, 2004.
- [8] Microchip Technology Inc., <http://www.microchip.com>.
- [9] RoboTech srl, <http://www.RoboTechsrl.com>.
- [10] mikroElektronika, www.mikroe.com.

Weight-Based Fair Rate Allocation in Resilient Packet Ring

Elyas Mohamadzadeh Kosari
Department of Computer
Ferdowsi University of Mashhad
#763, 55 Fallahi Street,
Fallahi Blvd., Mashhad, Iran
00985116638570
elyas.kosari@stu-mail.um.ac.ir

Mohammad Hossein Yaghmaee Moghaddam
Associate Professor
Department of Computer
Ferdowsi University of Mashhad
Computer Department, Faculty of engineering, Ferdowsi
University of Mashhad, Azadi Square
Mashhad, Iran
00985118830501
yaghmaee@ieee.org

Abstract— Among different Metro technologies, Rings are the most prevalent ones. There are different kinds of ring technologies such as SONET and Gigabit Ethernet. SONET rings suffer from low utilization and Gigabit Ethernet rings suffer from unfairness. Resilient Packet Ring is a technology that inherits good characteristics of both SONET and Gigabit Ethernet. It provides high utilization, spatial reuse and fairness. The current fairness algorithm of Resilient Packet Ring does not account station weights in calculating the fair rate. In this paper a weighted fairness algorithm will be proposed. Simulation results show the acceptable performance of the proposed algorithm.

Keywords—Resilient Packet Ring; Fairness; Congestion Control; Weighted Fairness; RIAS

I. INTRODUCTION

Rings are the most prevalent metro technologies because of their protection and fault tolerance properties, but the current metropolitan ring networking technologies exhibit several limitations [1]. A SONET ring can ensure the minimum bandwidth and hence fairness between any pairs of nodes. However, using circuits prohibits unused bandwidth to be reclaimed by other flows and results in low utilization. On the other hand, a Gigabit Ethernet (GigE) ring can provide full statistical multiplexing and high utilization, but suffers from unfairness. Finally, legacy ring technologies such as FDDI do not employ spatial reuse. In these technologies, a rotating token is used such that a node must have the token to be able to transmit; that is, only one node can transmit at a time [2].

The Resilient Packet Ring (RPR), standardized in IEEE 802.17 Workgroup, is a Medium Access Control (MAC) protocol for metro-ring networks [3]. RPR not only shares SONET's ability to provide fast recovery from link and node failure, but also benefits from Ethernet's cost and simplicity. RPR employs a dual-ring structure utilizing a pair of unidirectional counter-rotating rings, called ringlets [4], inner-ringlet is clockwise and outer-ringlet is counter-clockwise.

In RPR, packets are removed from the ring at the destination so that different segments of the ring can be used at the same time for different flows; as a result, the spatial reuse feature is achieved. Enabling the spatial reuse feature introduces the challenge of guaranteeing fairness among the nodes sharing the same link [1]. Therefore, it is essential that RPR provides a fairness and congestion control mechanism that not addressed by either SONET or Ethernet, which makes RPR to be significantly more efficient than other technology. The key performance objective of RPR is to simultaneously achieve high utilization, spatial reuse, and fairness [4].

The ring access scheme in RPR is based on the Buffer Insertion Ring (BIR) method, in which every node on the ring has one or more insertion buffer(s) [4], called transit buffer(s). The ring traffic transiting a node, may be temporarily stored in the transit buffer. Each node is allowed to add its local traffic to the ring when its transit buffer is empty. In other words, ring traffic has non-preemptive priority over the local traffic. This means that the transit traffic at each node can block the local traffic of that node in accessing the ring [3]. Hence, downstream nodes may suffer from starvation problem if upstream nodes keep bursting traffic [5]. To avoid this problem, all nodes should be forced to adjust the insertion rate of their local traffic according to their fair shares [3].

The objective of the fairness algorithm in RPR is to distribute fairly the available bandwidth on any link among the local traffic of all competing nodes on that link [3]. In [6] the *fair rate* is defined as the rate at which every node should add its local traffic to the ring without starving its downstream nodes. When a node is congested, it calculates the fair rate. The calculated fair rate is then advertised to the upstream nodes through a control message. As the upstream nodes receive the control message, they adjust their *add_rates* according to the received fair rate [3].

Improvements to the fairness algorithm of [6] are proposed in [7-11]. The weight aspect of the fairness algorithm has been investigated in [12]. In this paper we will focus on the weighted fairness in another way.

The rest of this paper is organized as follows. In Section II we will provide an overview of RPR node architecture and its fairness algorithm. In Section III the proposed weighted fairness algorithm will be discussed. Performance evaluation of the proposed algorithm is provided in Section IV. Finally, Section V concludes the paper.

II. OVERVIEW OF RPR

In this section, we describe the basic operations of RPR MAC. Please refer to [6] for detailed descriptions.

RPR supports three different service classes: high-priority guaranteed bandwidth, low-delay, low-jitter Class A; medium priority, low-jitter, bounded delay Class B; and best-effort Class C. Class B is divided to two subclasses: Class B Committed Information Rate (B-CIR) which has guaranteed bandwidth, and Class B Excess Information Rate (B-EIR) which is a best-effort service. Class B-EIR and Class C are called fairness-eligible traffic and can reclaim unused bandwidth [6]. In this paper we consider only fairness-eligible traffic.

Fig. 1 illustrates the RPR basic node architecture, where only the traffic of one of the ringlets is shown. As it is shown in Fig. 1, the transit buffer can be implemented in two modes [6]: Single Queue Mode (SQM) and Dual Queue mode (DQM). In SQM (Fig. 1-a) the transit path consists of a single FIFO queue, which is referred to as Primary Transit Queue (PTQ). In this mode the scheduler gives strict priority to transit traffic, i.e. PTQ over station traffic. In DQM (Fig. 1-b) there are two queues for transit path, Primary Transit Queue (PTQ) for guaranteed Class A traffic, and Secondary Transit Queue (STQ) for Class B and Class C traffic. In this mode the scheduler always services PTQ first. If this queue is empty the scheduler services Class A and Class B station traffic, respectively. After that the scheduler employs round robin service between Class C station traffic and STQ until a threshold on STQ is reached. If STQ reaches the buffer threshold, STQ traffic is always selected over station traffic to ensure lossless transit path [6].

In both modes the objectives are hardware simplicity and to ensure that the transit path is lossless, i.e. once a packet is injected to the ring, it will not be dropped at a downstream node [6]. We will utilize the dual queue mode in this paper.

The RPR standard defines a fairness algorithm that specifies how upstream traffic should be throttled according to downstream measurements, namely, how a congested node will send fairness messages upstream so that upstream nodes can appropriately configure their rate limiters to throttle the rate of injected traffic to the fair rate. The nodes that send over the same congested link are known in RPR standard [6] as a *congestion domain*. The node directly upstream of the most congested link is called the *head* of the congestion domain. The node in the congestion domain that is furthest away from the head is called the *tail* of the congestion domain.

The RPR fairness algorithm has two modes of operation: Aggressive Mode (AM) and Conservative Mode (CM). In both modes, when a downstream node is congested, it conveys its congestion state to upstream nodes so that they throttle their

traffic and ensure that there is sufficient capacity for the downstream traffic. To do so, the congested node calculates its *local_fair_rate* as the service rate of its own traffic, *add_rate*, and sends a fairness control message containing the calculated fair rate to the upstream nodes. Each upstream node that is sending to the congested link, must throttle its traffic according this same rate [6].

Aggressive mode is the default mode of operation of the RPR fairness algorithm. In this mode, node *n* is congested if one of the following conditions is satisfied [6]:

$$STQ_depth[n] > low_threshold \tag{1}$$

or

$$forward_rate[n] + add_rate[n] > unreserved_rate \tag{2}$$

Where, *STQ_depth* is the occupied amount of STQ, *low_threshold* is a fraction of STQ, *forward_rate* is the rate a node forwards traffic from STQ, *add_rate* is the rate a node adds its own traffic to the ring, and *unreserved_rate* is the link capacity minus the reserved rate for guaranteed traffic. In this paper we consider only fairness-eligible traffic; therefore, the *unreserved_rate* is the link capacity in the rest of this paper.

In conservative mode, node *n* is said to be congested if the following condition is satisfied [6]:

$$forward_rate[n] + add_rate[n] > low_threshold \tag{3}$$

Where, *low_threshold* is a rate-base parameter that is a fixed value less than the link capacity. In this paper we focus on conservative mode of operation.

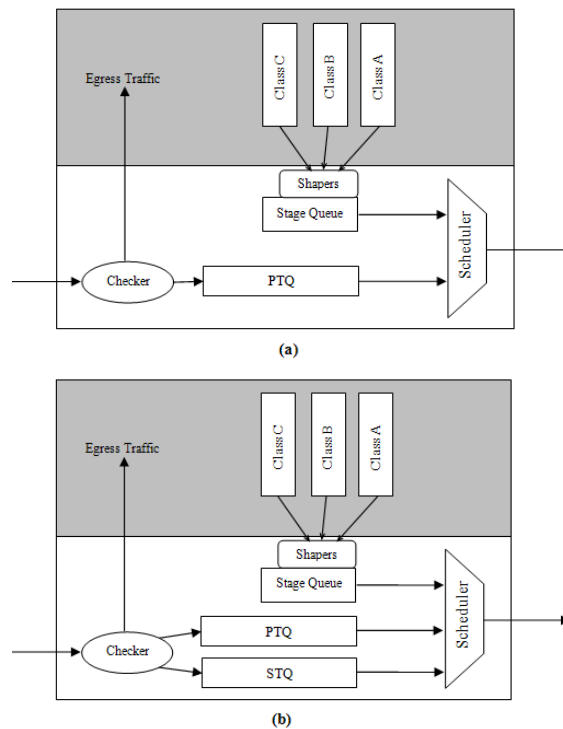


Figure 1. RPR Node Architectures: a) Single Queue Mode and (b) Dual Queue Mode

RPR fairness algorithm is based on ingress aggregation, which is referred to as "Ring Ingress Aggregated with Spatial Reuse (RIAS)" fairness [7]. RIAS fairness has two key components. The first component defines the level of traffic granularity for fairness determination at a link as an ingress-aggregated (IA) flow, i.e., the aggregate of all flows originating from a given ingress node, but not necessarily destined to a single egress node. RIAS reference model ensures that an ingress node's traffic receives an equal share of bandwidth on each link relative to other ingress nodes' traffic on that link. The second component of RIAS fairness ensures maximum spatial reuse subject to this first constraint. That is, bandwidth can be reclaimed by IA flows when it is unused either due to lack of demand or in cases of sufficient demand in which flows are bottlenecked elsewhere [7]. Detailed description of RIAS is provided in [7].

The RIAS fairness definition does not include the station weights in calculating the fair rate. In this paper, in order to provide a better fair rate calculation, a fairness algorithm with the inclusion of the administratively assigned weights to the stations will be proposed.

III. WEIGHTED FAIRNESS ALGORITHM

In this section, before we propose our fairness algorithm, we provide an example scenario to show the importance of considering weights in calculating the fair rate.

A. Example Scenario

A service provider offers Internet and video services over its RPR network using OC12 links (622 Mbps) as shown in Fig. 2. The provider wants to make sure that there is always enough bandwidth to accommodate the video requests of the subscribers. The video server is connected to Station 1 and the Internet connection is through Station 2 on the ring. The service provider is utilizing MPEG-2 compression for a high-definition video service where each connection is taking approximately 4 Mbps of bandwidth [13].

Assume that a total of 100 different channels are being requested by the customers of the video service, i.e. a total of 400 Mbps of traffic. These customers are connected to Stations 5 and 6 on the ring. At the same time, some 300 other customers with 1 Mbps Internet connection services at Stations 5 and 6 are downloading files through the Internet, generating a total traffic of 300 Mbps. In this scenario the inner-ringlet is used.

In RIAS fairness, that does not consider weights in calculating the fair rate, stations will share the ring bandwidth equally, i.e. each of Stations 1 and 2 will generate traffic at 311 Mbps, as shown in Fig. 3.

Therefore, the service provider will not be able to accommodate the requests for 100 different channels. In this scenario, only 77 different channels can be distributed.

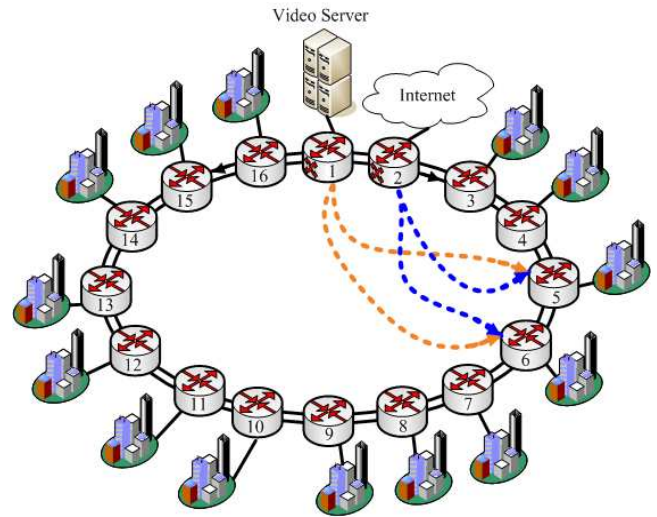


Figure 2. Scenario 1: Video and Internet Services

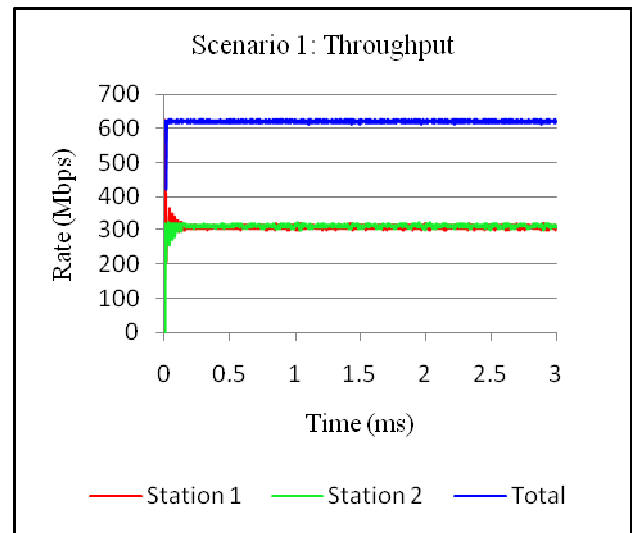


Figure 3. The Throughput of Scenario 1 without considering weights

B. Proposed Weighted Fairness Algorithm

In this section we will describe the proposed weighted fairness algorithm. As mentioned earlier in Section II, RPR fairness algorithm is based on RIAS fairness. In RIAS fairness, bandwidth is allocated to stations based on ingress aggregated flows, which is the aggregation of all the traffic originating from the same ingress node but not necessarily destined to the same station. In this fairness algorithm, when a node is congested it must calculate a local fair rate and adjust its *add_rate* based on this rate. Next, the congested station will demonstrate the calculated fair rate to upstream stations in a fairness control message. Each upstream node receiving this message will adjust its *add_rate* based on the received fair rate as well. The congested node calculates its local fair rate as in (4):

$$fair_rate = unreserved_rate / N \quad (4)$$

Where, *unreserved_rate* is the link capacity in this paper, since we only simulate fairness eligible traffic, and *N* is the number of active stations sending traffic through the congested link. As it is shown in (4), the problem in RIAS fairness is that the station weights are not considered in calculating fair rate.

To solve this issue we add station weights to the fair rate calculation. In the proposed weighted fairness algorithm, in case of congestion, the congested node, station *n*, will first calculate a normalized fair rate as follows:

$$\text{normalized_fair_rate} = (1 / \sum W_i) * \text{unreserved_rate} \quad (5)$$

Where, $\sum W_i$ is the sum of weights of active stations. It is used to normalize the rates.

Then, station *n* will calculate its *add_rate* as follows:

$$\text{add_rate}_n = w_n \times \text{normalized_fair_rate} \quad (6)$$

Where w_n is the weight of station *n*. Equation (6) is station *n*'s weighted fair share of the total unreserved bandwidth available. Station *n* is the head of the congestion domain.

The *normalized_fair_rate* value is also sent to upstream nodes. This message travels upstream through the ring, from head of the congestion domain until it reaches the tail of the congestion domain. Upstream stations that are in congestion domain and receive this rate will multiply their weight to this value to calculate their own weighted share of bandwidth, i.e. their *add_rate*. The *add_rate* calculation for station *i* is shown in (7):

$$\text{add_rate}_i = w_i \times \text{normalized_fair_rate} \quad (7)$$

As it is shown in (5) to (7), the station weights are now considered in the fair rate calculation and each node has its own weighted share of bandwidth. All the nodes in the congestion domain will send according to their fair, weighted share of bandwidth until the congestion period passes.

After the congestion period is passed, the nodes, which were in congestion domain, will increase their rates gradually according to (8):

$$\text{add_rate}_i = \text{add_rate}_i + \alpha \quad (8)$$

Where α is the ramp up coefficient, which is used to increase the rates. The rates are increased continually, until a new congestion occurs so that the new fair rate must be calculated. In the next section we will evaluate the proposed weighted fairness algorithm.

IV. SIMULATION RESULTS

In this Section performance evaluation of the proposed algorithm is presented. In order to demonstrate the performance evaluation we consider two different scenarios. The scenarios have been executed on the RPR simulator model developed at Simula Research Laboratory [14]. In both scenarios, STQ size is 512 KB, *lp_coef* of the RPR MAC is set to 16, and the link rate is 622 Mbps.

The *lp_coef* parameter is the low-pass filter coefficient. In [6], each rate is smoothed with respect to past rate measurements, periodically. This smoothing method is known as low-pass filtering. Low-pass filtering is done by setting a configurable parameter called *lp_coef*. In this paper we will set *lp_coef* to 16 (allowed values are 16, 32, 64, 128, 256 and 512), i.e. using a less smoothing low-pass filter.

We described that the fairness algorithm computes the fair rate every time congestion occurs on the ring. Furthermore, a station periodically re-computes its fair rate information. The period between computations is known as the aging interval. The aging interval value for the link rates greater than or equal to 622 Mbps is 100 μ s. By experimenting with the sample period, it seems that 30 aging intervals, i.e. 3 ms, is a reasonable sampling period. Using even shorter sampling periods gives visually the same results. Therefore, every scenario will be executed for 3 ms and the results will be shown in terms of charts.

The first scenario is the one shown in Fig. 2 and described in Section III-A. In this scenario, there are 16 stations on the ring. Stations 1 and 2 started injecting traffic to the ring at the time 0s. In this scenario we need 400 Mbps bandwidth to be able to service video customers at all times. In this case the remaining bandwidth is 622-400=222 Mbps and it should be allocated to Internet customers. Since the ratio between these rates is approximately 2, a weight of 2 is assigned to Station 1, while the weight of Station 2 remains 1. Fig. 4 shows the throughput of the scenario.

As it is illustrated in Fig. 4, the bandwidth is correctly allocated to Stations 1 and 2. The video server is assigned 400 Mbps of bandwidth and the remaining bandwidth, i.e. 222 Mbps is assigned to Internet service. In this way, all of the 100 video channels would be serviced completely. Fig. 4 shows the correct behavior of the proposed weighted fairness.

The second scenario is depicted in Fig. 5. There are 8 stations in this scenario. As it is shown in Fig. 5, an audio server is connected to Station 1, Internet is connected to Station 2, video server is connected to Station 3 and FTP server is connected to Station 4. The subscribers of the first three services above are connected to Station 5, and the subscribers of the FTP are connected to Station 6. The service provider is utilizing MPEG-2 compression, with 4 Mbps each connection, for video service, and high definition audio, with 3 Mbps each connection, for audio service.

There are 90 different channels requested from video service and also 40 another channels are requested from audio service. At the same time there are 100 customers each requesting 1 Mbps Internet service and another 100 customers requesting 1 Mbps FTP connections. The bandwidth requested for these services include 360 Mbps for video service, 120 Mbps for audio service, 100 Mbps for Internet and 100 Mbps for FTP service. The total requested bandwidth is 680 Mbps while the available bandwidth is 622 Mbps. As described before, this is a case of congestion.

If the provider wants to accommodate all requested bandwidth of audio and video services, he/she must allocate a total of 480 Mbps of bandwidth to these services. Thus, the

remaining bandwidth would be $622-480=142$ Mbps. Since FTP and Internet services have the same priority from the provider's point of view, the remaining bandwidth could be allocated to these services equally, i.e. approximately 71 Mbps of bandwidth each. Since the ratio between 120 (audio requested bandwidth) and 71 (FTP or Internet requested bandwidth) is approximately 2 we will assign a weight of 2 to audio and keep the FTP and Internet servers' weights to 1. With the same logic, the ratio between 120 and 360 (video requested bandwidth) is 3. Since we assigned a weight of 2 to audio service and $2 \times 3=6$, we will assign a weight of 6 to the video server.

The Throughput of the second scenario, Fig. 5, is illustrated in Fig. 6. All the services start injecting traffic to the ring at time 0s. As it is depicted in Fig. 6, the bandwidth is allocated to each of these services based on their assigned weights, correctly. Station 1, 2, 3, and 4 consume 120 Mbps, 71 Mbps, 360 Mbps, and 71 Mbps of bandwidth, respectively, which verifies the correct behavior of the proposed algorithm.

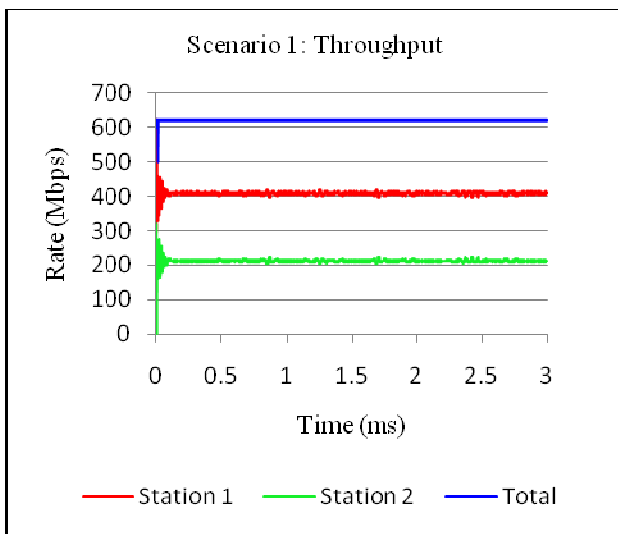


Figure 4. The Throughput of Scenario 1 with considering weights

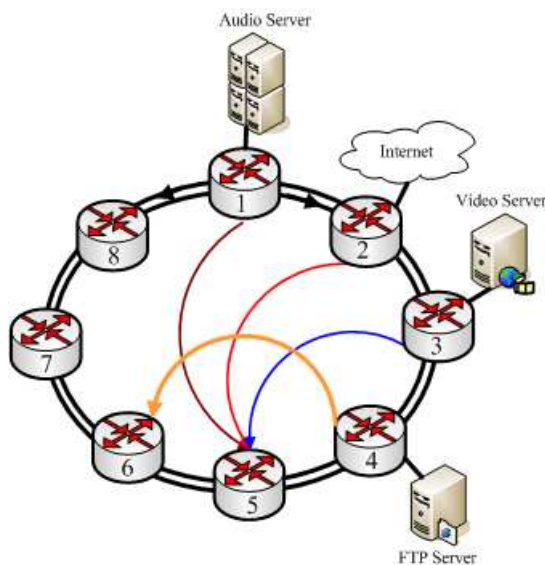


Figure 5. Scenario 2: Audio, Internet, Video, and FTP Services

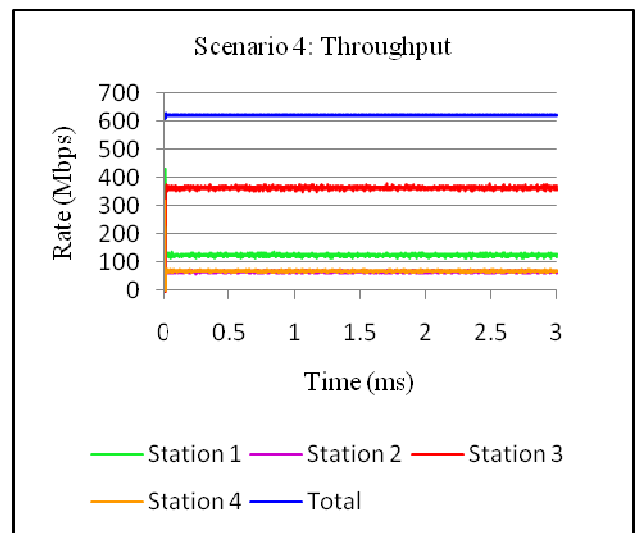


Figure 6. The Throughput of Scenario 2 with considering weights

V. CONCLUSION

Rings are the most popular topologies because of their cost and fault tolerance properties. Among ring technologies, RPR is one of the most popular ones, because it provides high utilization, spatial reuse and fairness simultaneously.

One of the most important challenges of RPR is to provide fairness among stations. The current RPR fairness algorithm is based on RIAS fairness, which does not consider weights of stations in calculating the fair rate.

In this paper a weighted fairness algorithm was proposed. We have simulated different scenarios to show the efficiency of our proposed algorithm. Simulation results show the good and acceptable performance of the proposed algorithm.

REFERENCES

- [1] F. Alharbi, N. Ansari, "Low complexity distributed bandwidth allocation for resilient packet ring networks," *High Performance Switching and Routing, 2004. HPSR. 2004 Workshop on*, vol., no., pp. 277- 281, 2004
- [2] P. Yue, Z. Liu; J. Liu, "High performance fair bandwidth allocation algorithm for resilient packet ring," *Advanced Information Networking and Applications, 2003. AINA 2003. 17th International Conference on*, vol., no., pp. 415- 420, 27-29 March 2003
- [3] A. Shokrani, I. Lambadaris, J. Talim, "On modeling of fair rate calculation in resilient packet rings," *Computers and Communications, 2005. ISCC 2005. Proceedings. 10th IEEE Symposium on*, vol., no., pp. 651- 657, 27-30 June 2005
- [4] X. Zhou, G. Shi, H. Fang, L. Zeng, "Fairness algorithm analysis in resilient packet ring," *Communication Technology Proceedings, 2003. ICCT 2003. International Conference on*, vol.1, no., pp. 622- 624 vol.1, 9-11 April 2003
- [5] C. Huang, H. Peng, F. Yuan, J. Hawkins, "A steady state bound for resilient packet rings," *Global Telecommunications Conference, 2003. GLOBECOM '03. IEEE*, vol.7, no., pp. 4054- 4058 vol.7, 1-5 Dec. 2003
- [6] IEEE Std. 802.17: Resilient packet ring (RPR) access method and physical layer specification, September 2011
- [7] V. Gambiroza, P. Yuan, L. Balzano, Y. Liu, S. Sheafor, and E. Knightly, "Design, analysis, and implementation of DVSR: a fair high-performance protocol for packet rings," *IEEE/ACM Trans. Networking*, vol. 12, no. 1, pp. 85-102, 2004.

- [8] F. Davik, A. Kvalbein, S. Gjessing, "Performance evaluation and improvement of non-stable resilient packet ring behavior," in *Proc. Part II of the 4th Int. Conf. on Networking (ICN'05)*, LNCS 3421, Reunion Island, 2005, pp. 551–563.
- [9] F. Alharbi, N. Ansari, "Distributed bandwidth allocation for resilient packet ring networks," *Comput. Netw.*, vol. 49, no. 2, pp. 161–171, Oct. 2005.
- [10] F. Alharbi, N. Ansari, "SSA: simple scheduling algorithm for resilient packet ring networks," *IEE Proc.-Commun.*, vol. 153, no. 2, pp. 183–188, Apr. 2006.
- [11] F. Alharbi, N. Ansari, "Allocating bandwidth in the resilient packet ring networks by PI controller", in Proc. of Sarnoff Symposium, May 2011, pp. 1 - 6
- [12] M. Yilmaz, N. Ansari, "Weighted Fairness and Correct Sizing of the Secondary Transit Queue in Resilient Packet Rings," *Optical Communications and Networking, IEEE/OSA Journal of*, vol.2, no.11, pp.944-951, November 2010
- [13] S. Kwon, A. Tamhankar, K. R. Rao, "Overview of H.264/MPEG-4 part 10", *Journal of Visual Communication and Image Representation*, Volume 17, Issue 2, pp. 186-216, 2006
- [14] "Java based RPR simulator" Simula Research Laboratory, Available at: <http://simula.no/research/nd/software/rpr>.

Data Processing in FPGA-based Systems

Valery Sklyarov
University of Aveiro
Department of Electronics, Telecommunications and
Informatics/IEETA
Aveiro, Portugal

Iouliia Skliarova
University of Aveiro
Department of Electronics, Telecommunications and
Informatics/IEETA
Aveiro, Portugal

Abstract—More than 50% of FPGA market is in the scope of communication and information. Data processing is a very broad area that is important for numerous computations, networking, embedded systems, Internet-based applications, etc. Many problems in this area are computationally intensive and thus, they require parallelization and acceleration based on new technologies. FPGAs can be seen as a very attractive platform permitting application-specific software and problem-targeted hardware to be combined on a single configurable microchip. The tutorial is dedicated to hardware-oriented data processing algorithms with emphasis on parallelism and pipelining. FPGA-targeted design flow is based on the rational use of configurable logic blocks/logic elements and embedded components (DSP slices and block RAMs). Particular examples, potential practical applications, experiments and comparisons will be demonstrated.

Index Terms—Data processing, high-performance computing, parallelism, pipelining, FPGA.

I. INTRODUCTION

Data processing plays essential role in vast variety of computational systems used in the scope of information and communication. Such tasks as sorting, searching, frequent items computations and their varieties are the most requested [1-3]. For large volumes of data these tasks are time consuming and speed-up is greatly required. A number of recent research works are targeted to the potential of advanced hardware accelerators, which are analyzed in detail in [3-6]. Notable results have been achieved through applying parallelism, pipelining, non-sequential circuits, and other techniques and building specialized blocks in hardware. A special attention has been paid to such competitive implementation platforms as: field-programmable gate arrays – FPGA (ex. [3,4,7-12]), graphics processing units – GPU (ex. [5,8,13-15]), and multi-core central processing units – CPU (ex. [16,17]).

One of the most important features of FPGA-based circuits is an opportunity to build the entire system that is composed of various components (data processing blocks can be among them). Optimization of resources for any component permits the same microchip to be used for implementing additional tasks. High performance is important for many practical applications. Significant speed-up can be achieved if hardware circuits are used as accelerators for general-purpose and application-specific software. A number of comparisons (FPGA vs. multi-core CPU, FPGA vs. GPU, FPGA vs. DSP) can be found in [3,10,16,18,19]. As a general conclusion, we can say that FPGAs become beneficial for more and more

cases mainly due to inherent configurability and relatively cheap development cost. Generally, FPGA-based circuits are more redundant and slow compared to CPU and GPU. However, they allow easier to create operations and blocks that are indeed required for particular applications. For example, in parallel sorting, the size of operands can be customized and we can combine reasonably combinational and sequential circuits. Besides, multi-core CPU and hardware accelerators can be implemented on the same microchip (see, for instance, Zynq 7000 from Xilinx [20]). Our experience has shown that the best FPGA-based implementations should be as much regular as possible simplifying interconnections and the routing procedure. Thus, achieving the highest possible regularity of circuits is the main priority of this tutorial.

II. PROCESSING SMALL SETS OF DATA

Answering the question whether a set of data is *large* or *small* depends on many factors. A set that was considered to be *large* some time ago is qualified as *small* today. We will say that a set is *small* if all data items can be processed in parallel on modern microchips or computers. A good model of a parallel circuit is a *sorting network* composed of *comparators* [1]. Different types of such networks and their implementations in hardware were widely investigated [4,5,6,9,10]. Fig. 1 gives a simple example of data sort with the aid of an *odd-even transition network* [1,6] which is composed of vertical lines of comparators (numbered at the top) and each comparator can be described in VHDL as follows (see also Fig. 1c and 1d):

```
MaxValue <= Op1 when Op1 >= Op2 else Op2;  
MinValue <= Op2 when Op1 >= Op2 else Op1;
```

Given data (66,82,80,102,37,138,40,150) are sorted in descending (Fig. 1a) and ascending (Fig. 1b) order. Each vertical line is composed of $N/2$ or $N/2-1$ comparators (N is the number of data items) and there are totally $C(N)=N \times (N-1)/2$ comparators. Note that if there is one additional comparator which processes items with indices 0 and $N-1$ at the network inputs then one vertical line can be removed from Fig. 1.

If data items are swapped they are shown in *italic* and underlined. If there is no exchange in any pair of lines then all data are sorted. Hence, the decision about the result can be taken earlier than after propagation through all the lines.

The circuits in Fig. 1a and Fig. 1b might be combinational and sequential. Sequential circuit can be represented in a way shown in Fig. 2. Fig. 2a includes two blocks: a *register* and *comparators*. Fig. 2b connects the circuit in Fig. 1a in a

pipeline. Complexities C_{1ab} , C_{2a} and C_{2b} of the circuits in Fig. 1a,b, Fig. 2a and Fig. 2b are different: $C_{1ab} = C(N)$ comparators; $C_{2a} = N/2$ or $N-1$ or more comparators plus N -item register; $C_{2b} = k \times (N/2$ or $N-1$ or more comparators) plus k N -item registers, where k is the number of steps in the pipeline. Draft evaluation of different circuits permits easily to conclude that Fig. 2b represents the fastest and the most complicated circuit and Fig. 2a represents the least complicated circuit. The circuits in Fig. 1a,b are combinational and they have an excessive delay from inputs to outputs. The total delay is equal to $N \times t$, where t is a delay of each vertical line. If the circuit in Fig. 2a generates a special *enable* signal, which is equal to zero at any iteration when there is no swap (exchange) of data, then we can get delay equal to $2 \times t$ (one even and one odd iteration) in the best case. Indeed, suppose we need to sort data that occasionally have been received in the sorted order, let us say: 8,7,6,5,4,3,2,1. The sequential circuit (Fig. 2a) can conclude that data have already been sorted in time $2 \times t$ and the combinational circuit in Fig. 1a still needs time $N \times t$.

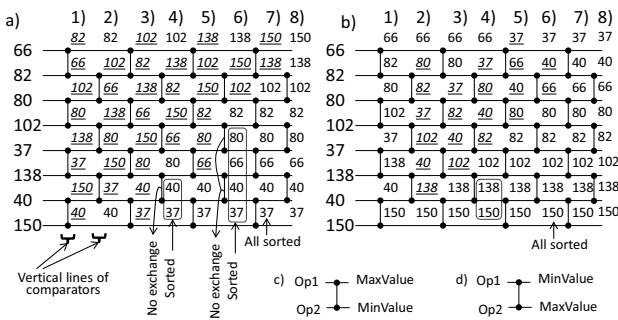


Fig. 1. Simple example of the odd-even transition sorting network.

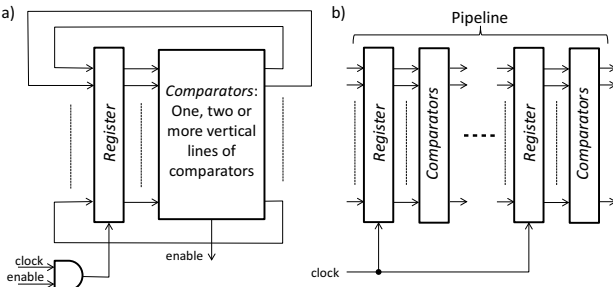


Fig. 2. Sequential circuits representing sorting networks.

It is known [6] that the *odd-even transition network* is not ideal in both resource and time consumption. For example an *odd-even merge sorting network* [21] is considered to be significantly more efficient and it was even characterized in [6] as almost optimal. Fig. 3 depicts such network for our example.

The network in Fig. 3 has smaller depth (6 instead of 8 in Fig. 1) and less comparators (19 instead of 28). However, we cannot conclude that data are sorted until reaching the last vertical line of comparators and, thus, sorting of data 1,2,3,4,5,6,7,8 still needs time equal to $D(N) \times t$, where $D(N)$ is the depth [1,4] of the network (minimal number of steps that have to be executed; in Fig. 3 $D(N)=6$). Besides, sequential

circuits for *odd-even merge sorting networks* (like shown in Fig. 2a) would need complex interconnections between lines through multiplexers which increase resources and propagation delay.

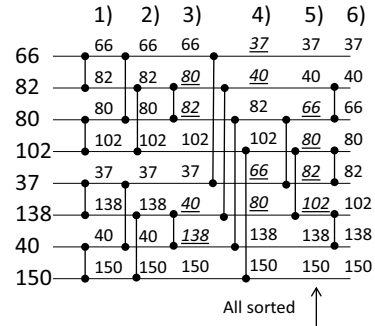


Fig. 3. The odd-even merge sorting network for our example.

It is known that a *bitonic merge network* is one of the fastest [1,5,22]. One possible representation of such network [1] is shown in Fig. 4 for our example in a non-standard form because comparators have different types. Any arrow points to greater or equal item after comparison and eventual swapping.

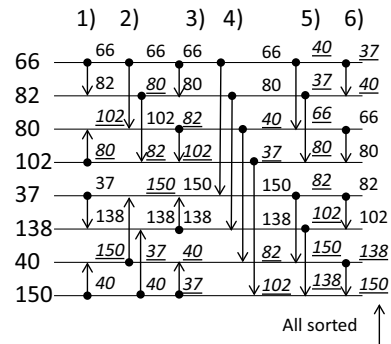


Fig. 4. A non-standard network based on bitonic sorting for our example.

The circuit in Fig. 4 has the same problems as the circuit in Fig. 3 and besides it is composed of different comparators making hardware description heterogeneous.

The tutorial will analyze and evaluate different types of networks (some of which are shown in Fig. 1-4). We found that the circuit in Fig. 2a gives a good compromise between complexity and performance. Pure combinational solutions have many limitations. For example, for circuits in Fig. 1a,b resources of the Xilinx Spartan-6 FPGA XC6SLX45 are not sufficient for $N=32$ (for $M=32$ bit data items). The results of [4] show that even for the more advanced FPGA XC5VFX130t of Xilinx Virtex-5 family $N \leq 64$ ($M=32$). On the other hand the circuit in Fig. 2a can be implemented in the indicated above Spartan-6 FPGA for $N=256$ ($M=32$). If we compare two circuits shown in Fig. 1a and Fig. 2a for $N=M=32$ we can see that their performance is almost the same while hardware resources for the circuit in Fig. 2a are decreased in more than 6 times. Table I presents the number of comparators $C(N)$ and the depth $D(N)$ of different sorting networks for N data items [1,4,22]. For sizes $N=2^p$ the exponent p is given.

TABLE I. NUMBER OF COMPARATORS C(N) AND DEPTH D(N) OF DIFFERENT NETWORKS

bubble/insertion	odd-even transition	odd-even merge	bitonic merge
$C(N)=N \times (N-1)/2$; $D(N)=2 \times N-3$	$C(N)=N \times (N-1)/2$; $D(N)=N$	$C(N=2^p)=(p^2-p+4) \times 2^{p-2}-1$; $D(N=2^p)=p \times (p+1)/2$	$C(N=2^p)=(p^2+p) \times 2^{p-2}$; $D(N=2^p)=p \times (p+1)/2$

III. PROCESSING LARGE SETS OF DATA

Very often large sorted data sets are built from small sorted data sets applying different types of merging (see, for instance, [9,11]). Fig. 5 gives an example, where four preliminary sorted sub-sets are combined in one sorted set. Items from sub-sets are transmitted to the merging circuit sequentially beginning from the largest items. Four items are compared at each step and an item with the maximal value is dispatched to the output and is replaced with a new item from the same sub-set (see Fig. 5a). Any sub-set is blocked if it is empty (see Fig. 5b showing sub-sets after the first three steps). Thus, the time of merging is equal to $N \times t_m$, where N is the total number of items in all sub-sets and t_m is time needed for each step. Using dual port memories permits to sort data in two directions: one direction for maximal values and one direction for minimal values. Hence, sorting can be executed twice faster.

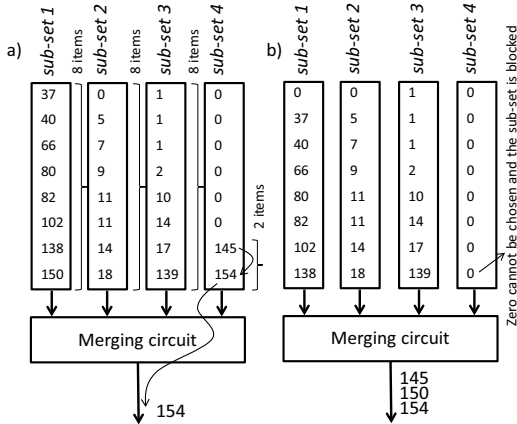


Fig. 5. Merging of sorted sub-sets.

There is another efficient way to sort large sets of data called address-based sorting [12] which has similarity with *pigeonhole (count)*, *basket sort*, *counting sort*, and *distribution sort*. The main idea is rather simple. As soon as a new data item is taken, its value V is considered to be the address of memory to record a flag (1). We assume that memory is zero filled first. Fig. 6 gives an example.

As soon as all input data are recorded in memory in the form of a long binary vector $0...010...010...010...010...010...010...01...01...$ shown in Fig. 6, the sorted sequence can be produced immediately. The vector above has to be converted to item values and this can easily be done (see [12] for details).

There are some problems with this kind of sort:

- Memory size is very large and equal to 2^M ,
- If $N \ll 2^M$ then memory includes many empty holes.
- Data items cannot be repeated.

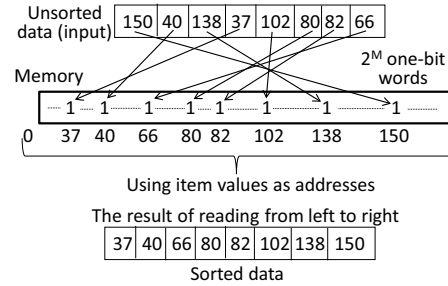


Fig. 6. The core of the address-based data sort.

However, we found that all these problems can efficiently be solved applying hierarchical structuring, merging and sorting networks described above. The tutorial will present clear explanation of these ideas.

IV. PARTICULARITIES OF DATA PROCESSING IN FPGA

Modern FPGAs include configurable logic blocks (see Table II, where DRAM is distributed RAM) and numerous embedded components which can effectively be used for data processing. For example, embedded to Xilinx FPGAs DSP slices contain 48-bit adder with accumulator and can be employed for comparisons. The following VHDL code demonstrates how one DSP48A1 slice [23] (available for Xilinx Spartan-6 FPGA family) permits two pairs of 16-bit operands (D,B) and (C,A) to be compared (the sequence D,B and C,A is chosen to provide direct association with names of inputs and outputs in [23]):

```

C <= D(11 downto 0) & C & "11111111111111111111";
D_greater_B <= Result(17); -- Result is P output of DSP slice
C_greater_A <= not Result(47);
maxD_B <= D when D_greater_B = '1' else B;
maxC_A <= C18 when C_greater_A = '1' else A;
DSP48A1_inst : DSP48A1 -- OPMODE => "11011111"

```

OPMODE selects the operation assigning to the not accumulated output P the following value: $C-(D:A:(D-B))$ (see details in [23]). Thus, comparison invokes a subtraction operation. This slice can also be used for comparison of two 48 bit operands. More advanced FPGA of Xilinx Virtex-5, Virtex-6 and Virtex-7 families include DSP48E and DSP48E1 with extended functionality and they can be used even more efficiently. It should be noted that DSP slices are directly targeted to digital signal processing (not to data processing). Nevertheless, an analysis of these slices clearly demonstrates that they have all necessary functionality for the required operations over data. The only problem is the lack of some outputs. We believe that adding such outputs would permit to consider DSP slices as a good alternative and amazing competitor to GPU for stream processing in SIMD (single instruction multiple data) mode. Table III shows the number of embedded DSP slices for different families of Xilinx FPGAs.

Another powerful feature of FPGA is availability of embedded dual-port block RAMs with configurable ports. For example, Xilinx FPGAs include 36 Kb (18 Kb) blocks, which can be configured as: $32K \times 1$, $8K \times 4$, $4K \times 9$, $2K \times 18$, $1K \times 36$, or 512×72 ($16K \times 1$, $8K \times 2$, $4K \times 4$, $2K \times 9$, $1K \times 18$, or 512×32)

RAMs. Besides, it is allowed for two ports of the same block to be configured differently. Such blocks can efficiently be used for fast parallel operations over multi-item vectors, which is very helpful for different sorting networks. Since interaction with memory can be done using ports with different configurations (e.g. 32K×1 for writing and 512×72 for reading), the address-based data sort can be structured easier and executed faster. Table III shows the number of embedded memory (in blocks and in kilobits - Kb) for different families of Xilinx FPGAs.

The tutorial will present detailed examples demonstrating the use of FPGA hardware in data processing.

TABLE II. SUMMARY OF CONFIGURABLE COMPONENTS FOR XILINX FPGAS

FPGA family	Logic cells	Configurable logic blocks		
		Slices	Flip-flops	DRAM (Kb)
Spartan-6 [24]	3840-147443	600-23038	4800-184304**	75-1355
Virtex-5 [25]	n/a	3120-51840	*	210-4200
Virtex-6 [26]	74496-758784	11640-118560	**	1045-8280
Artix-7 [27]	101440-360000	15850-56250	**	1188-4638
Kintex-7 [27]	65600-477760	10250-74650	**	838-6788
Virtex-7 [27]	284000-1954560	44375-305400	**	4388-21550
Zynq EPP [28]	28000-350000 ¹	4400-54650**	35200-437200**	***

* each slice contains 4 look-up tables (LUTs) and 4 flip-flops; ** each slice contains 4 LUTs and 8 flip-flops; *** 25-50% of slices can use their LUTs as distributed 64-bit RAM; 1 – only for reconfigurable logic in FPGA fabric

TABLE III. SUMMARY OF SOME EMBEDDED TO XILINX FPGA COMPONENTS

FPGA family	Embedded DSP slices	Embedded block RAM	
		in Kb	in blocks
Spartan-6 [24]	8-180*	216-4824	18Kb: 12-268
Virtex-5 [25]	24-1056**	936-18576	18Kb: 52-1032 36Kb: 26-516
Virtex-6 [26]	288-2016***	5616-38304	18Kb: 312-2128 36Kb: 156-1064
Artix-7 [27]	240-1040***	4860-18540	18Kb: 270-1030 36Kb: 135-515
Kintex-7 [27]	240-1920***	4860-34380	18Kb: 270-1910 36 Kb: 135-955
Virtex-7 [27]	840-3600***	16920-67680	18 Kb: 940-3760 36 Kb: 470-1880
Zynq EPP [28]	58-1080***	1920-17440	36K: 60-545

* DSP48A1; ** DSP48E; *** DSP48E1

The recently appeared extended processing platforms (EPP) [20,28] combine on the same microchip the high performance dual ARM® Cortex™-A9 MPCore™ with surrounding FPGA logic (Artix-7 or Kintex-7 FPGA). Thus, data processing can be done partially in software and partially in reconfigurable hardware.

V. OTHER METHODS FOR DATA PROCESSING IN FPGA

This section presents a brief review of other methods that have been used for FPGA-based data processing and are reported in numerous publications.

One of the most referenced simple algorithms is *bubble sort* (see Table I). Although this algorithm is rarely used in FPGA-based processing, one improvement named *comb sort* [29] was implemented in [9] and gave good results. The main idea of comb sort is to eliminate small values at the end of the list with data items because such items slow the *bubble sort* significantly.

One of the most frequently used methods in FPGA and GPU is *quicksort* (see, for example, [15]) which is a divide and conquer algorithm relying on a partition operation that divides a given set in sub-sets containing items smaller and larger than a specially chosen element called *pivot*. The first and the second sub-sets are then recursively sorted.

Radix sort is another frequently used non-comparison-based efficient algorithm that sorts numbers by processing individual digits. Effectiveness of this algorithm is shown in [5] comparing with the proposed in [5] special kind of bitonic sorters. The tutorial suggests *quicksort* and *radix sort* to be combined with the proposed methods and circuits offering very good results.

Different types of sorting over n-ary trees ($n \geq 2$) were considered for sequential [30] and parallel [31] implementations. This method can be efficient if it is combined with the address-based technique [12]. Using even the pure address-based technique in [32] demonstrates that very complex data sort problems (up to 4 billion of 32-bit data items) can be rapidly solved in a Virtex-6 FPGA with external 4 Gb DDR3 memory and these results can be further improved. Algorithms of sorting over trees for $n=2$ with hierarchical finite state machines are described in [33].

Less frequently used algorithms, such as *selection sort*, *insertion sort* (see Table I), *shell sort*, *heapsort*, will also be briefly analyzed and evaluated in the tutorial.

Many other types of data processing, such as that are needed for solving combinatorial search problems [34,35] as well as for implementing operations over Boolean and ternary matrices [36,37] will also be discussed.

VI. CONCLUSION

The tutorial will present in detail the topics briefly described in the previous sections with demonstration of the results from implemented in FPGA systems executing different data processing algorithms.

ACKNOWLEDGMENT

This research was supported by FEDER through the Operational Program Competitiveness Factors - COMPETE and by National Funds through FCT - Foundation for Science and Technology in the context of project FCOMP-01-0124-FEDER-022682 (FCT reference PEst-C/EEI/UI0127/2011).

REFERENCES

- [1] D.E. Knuth, The Art of Computer Programming. Sorting and Searching, vol. III. Addison-Wesley, 1973.
- [2] T.H. Cormen, C.E. Leiserson, R.L. Rivest, C. Stain, Introduction to Algorithms, 2nd edition. MIT Press, 2002.

- [3] R. Mueller, J. Teubner, and G. Alonso, "Frequent Item Computation on a Chip", *IEEE Trans. Knowledge and Data Engineering*, vol. 23, no. 8, 2011, pp. 1169-1181.
- [4] R. Mueller, J. Teubner, and G. Alonso, "Sorting Networks on FPGAs", *The International Journal on Very Large Data Bases*, vol. 21, no. 1, 2012, pp. 1-23.
- [5] G. Gapannini, F. Silvestri, and R. Baraglia, "Sorting on GPU for large scale datasets: A through comparison", *Information Processing and Management*, vol. 48, no. 5, 2012, pp. 903-917.
- [6] GPU Gems, Improved GPU Sorting, available at http://http.developer.nvidia.com/GPUGems2/gpugems2_chapter46.html.
- [7] D.J. Greaves and S. Singh, "Kiwi: Synthesis of FPGA circuits from parallel programs", *Proc. 16th IEEE Int. Symposium on Field-Programmable Custom Computing Machines - FCCM'08 USA*, 2008, pp. 3-12.
- [8] S. Chey, J. Liz, J.W. Sheaffery, K. Skadrony, and J. Lach, "Accelerating Compute-Intensive Applications with GPUs and FPGAs", *Proc. Symposium on Application Specific Processors - SASP'08, USA*, 2008, pp. 101-107.
- [9] R.D. Chamberlain and N. Ganesan, "Sorting on Architecturally Diverse Computer Systems", *Proc. 3rd Int. Workshop on High-Performance Reconfigurable Computing Technology and Applications - HPRCTA'09, USA*, 2009, pp. 39-46.
- [10] R. Mueller, *Data Stream Processing on Embedded Devices*. Ph.D. thesis, ETH, Zurich, 2010.
- [11] D. Koch and J. Torresen, "FPGASort: a high performance sorting architecture exploiting run-time reconfiguration on FPGAs for large problem sorting", *Proc. 19th ACM/SIGDA Int. Symposium on Field Programmable Gate Arrays - FPGA'11, USA*, 2011, pp. 45-54.
- [12] V. Sklyarov, I. Skliarova, D. Mihhailov, and A. Sudnitson, "Implementation in FPGA of Address-based Data Sorting", *Proc. 21st Int. Conf. on Field-Programmable Logic and Applications - FPL'11, Greece*, 2011, pp. 405-410.
- [13] X. Ye, D. Fan, W. Lin, N. Yuan, and P. Jenne, "High Performance Comparison-Based Sorting Algorithm on Many-Core GPUs", *Proc. IEEE Int. Symposium on Parallel & Distributed Processing - IPDPS'10, USA*, 2010, pp. 1-10.
- [14] N. Satish, M. Harris, and M. Garland, "Designing efficient sorting algorithms for manycore GPUs", *Proc. IEEE Int. Symposium on Parallel & Distributed Processing - IPDPS'09, Italy*, 2009, pp. 1-10.
- [15] D. Cederman and P. Tsigas, "A practical quicksort algorithm for graphics processors", *Proc. 16th Annual European Symposium on Algorithms - ESA'08, Germany*, 2008, pp. 246-258.
- [16] C. Grozea, Z. Bankovic, and P. Laskov, "FPGA vs. Multi-Core CPUs vs. GPUs", in: *Facing the multicore-challenge*, R. Keller, D. Kramer, and J.P. Weiss (Eds), Springer-Verlag Berlin, Heidelberg, 2010, pp. 105-117.
- [17] M. Edahiro, "Parallelizing fundamental algorithms such as sorting on multi-core processors for EDA acceleration", *Proc. of 18th Asia and South Pacific Design Automation Conf. - ASP-DAC'09, Japan*, 2009, pp. 230-233.
- [18] B. Cope, P.Y.K. Cheung, W. Luk, and L. Howes, "Performance Comparison of Graphics Processors to Reconfigurable Logic: A Case Study", *IEEE Transactions on Computers*, vol. 59, no. 4, 2010, pp. 433-448.
- [19] J. Gonzalez and R.C. Núñez, "LAPACKrc: Fast linear algebra kernels/solvers for FPGA accelerators", *Journal of Physics: Conference Series* 180, no. 1, 2009.
- [20] Xilinx, Documentation for Zynq-7000 EPP, available at: <http://www.xilinx.com/support/documentation/zynq-7000.htm>.
- [21] R. Sedgewick, *Algorithms in C++*, 3rd ed., Parts 1-4. Addison-Wesley, 1998.
- [22] K.E. Batcher, "Sorting networks and their applications", *Proc. AFIPS Spring Joint Computer Conf., USA*, 1968, pp. 307-314.
- [23] Xilinx, Spartan-6 FPGA DSP48A1 Slice User Guide, available at: http://www.xilinx.com/support/documentation/user_guides/ug389.pdf.
- [24] Xilinx, Spartan-6 Family Overview, 2011, available at: http://www.xilinx.com/support/documentation/data_sheets/ds160.pdf.
- [25] Xilinx, Virtex-5 Family Overview, 2009, available at: http://www.xilinx.com/support/documentation/data_sheets/ds100.pdf.
- [26] Xilinx, Virtex-6 Family Overview, 2012, available at: http://www.xilinx.com/support/documentation/data_sheets/ds150.pdf.
- [27] Xilinx, 7 Series FPGAs Overview, 2012 available at: http://www.xilinx.com/support/documentation/data_sheets/ds180_7Series_Overview.pdf.
- [28] Xilinx, Zynq-7000 All Programmable SoC Overview, 2012, available at: http://www.xilinx.com/support/documentation/data_sheets/ds190-Zynq-7000-Overview.pdf.
- [29] S. Lacey and R. Box, "A Fast, Easy Sort: A novel enhancement makes a bubble sort into one of the fastest sorting routines", *Byte*, vol. 16, no. 4, Apr., 1991, pp. 315-316, 318, 320.
- [30] V. Sklyarov, I. Skliarova, D. Mihhailov, and A. Sudnitson, "Processing Tree-like Data Structures for Sorting and Managing Priorities", *Proc. of IEEE Symp. on Comp. & Inf. - ISCI'11, Malaysia*, 2011, pp. 322-327.
- [31] D. Mihhailov, V. Sklyarov, I. Skliarova, and A. Sudnitson, "Parallel FPGA-based Implementation of Recursive Sorting Algorithms", *Proc. Int. Conf. on ReConfigurable Computing and FPGAs - ReConFig'10, Mexico*, 2010, pp. 121-126.
- [32] D. Mihhailov, A. Sudnitson, V. Sklyarov, and I. Skliarova, "Implementation of Address-Based Data Sorting on Different FPGA Platforms", *Proc. of 10th East-West Design & Test Symposium - EWDTs'12, Ukraine*, 2012.
- [33] I. Skliarova, V. Sklyarov, and A. Sudnitson, *Design of FPGA-based Circuits using Hierarchical Finite State Machines*. TUT Press, 2012.
- [34] J.D. Davis, Z. Tan, F. Yu, and L. Zhang, "A practical reconfigurable hardware accelerator for Boolean satisfiability solvers", *Proc. 45th ACM/IEEE Design Automation Conf. - DAC'08, USA*, 2008, pp. 780-785.
- [35] I. Skliarova and A.B. Ferrari, "Reconfigurable Hardware SAT Solvers: A Survey of Systems", *IEEE Trans. on Computers*, vol. 53, no. 11, 2004, pp. 1449-1461.
- [36] I. Skliarova and V. Sklyarov, "Design Methods for FPGA-based implementation of combinatorial Search Algorithms", *Proc. Int. Workshop on SoC and MCSoc Design - IWSOC'2006, 4th Int. Conf. on Advances in Mobile Computing and Multimedia - MoMM'2006, Indonesia*, 2006, pp. 359-368.
- [37] V. Sklyarov and I. Skliarova, "Architecture of a Reconfigurable Processor for Implementing Search Algorithms over Discrete Matrices", *Proc. Int. Conf. on Engineering of Reconfigurable Systems and Algorithms - ERSA'03, T.P. Plaks (ed.), USA*, 2003, pp. 127-133.

SESSION 3

ICT in Business Administration, Finance
and Economy

AICT 2012

Genetic Algorithm Approach for the Prediction of Business Risks' Dynamics of Enterprise

Gia Sirbiladze
Department of Computer Sciences
Iv.Javakhishvili Tbilisi State University
Tbilisi, Georgia
gia.sirbiladze@tsu.ge

Mikheil Kapanadze
Department of Computer Sciences
Iv.Javakhishvili Tbilisi State University
Tbilisi, Georgia
mikheil@mikheil.com

Abstract—This work deals with the problem of identification and modeling of Discrete Fuzzy Dynamic System (DFDS) with possibility uncertainty, using the technologies of Genetic Algorithms (GA). Applying the results from [5-9, 11-13,15,16,18-20], the fuzzy recurrent process, the source of which is expert knowledge reflections on the states of the evolutionary complex system, is constructed. The dynamics of DFDS is described and the constructed model is converted to the finite model. The DFDS transition operator is restored by means of expert data with possibility uncertainty. Obtained results are illustrated by the example for prediction and stopping problems for evaluations of the increasing business risks of the enterprise.

Keywords - DFDS; identification of DFDS; genetic algorithm; business risks.

I. INTRODUCTION

In recent years, both the dynamics of fuzzy system and its modeling problem have attracted significant attention. An application of the fuzzy dynamical systems in the prediction and stopping problems of complex business processes it isn't novel idea. Evidence exists that fuzzy models can explain cooperative processes, such as business and economics. Relationships between dynamics of fuzzy systems and the performance of decision support systems were found, and chaotic processes in various classes of fuzzy systems were proved as a powerful tool in analyzing complex systems with subjective uncertainty, as abnormal and monotone business processes.

In alternative classical approaches of modeling of the complex business processes the main accent is placed on the assumption of fuzziness. In such situations an exact quantitative analysis of real investigated business systems is apt not to be quite plausible. Hence the system approach to constructing models of complex systems with fuzzy uncertainty guarantees the creation of computer-aided systems forming the instrumental basis of the intelligent technology solutions of expert-analytic problems. It is obvious that the source of fuzzy-statistical samples is the population of fuzzy characteristics of expert knowledge. So, in this case the main difficulty is expert knowledge representation and management in the modeling of decision-making processes. For the construction of decision-making process to be more effective in the framework of computer systems that support this

process, we must solve analytical problems of optimization, state evaluation, model identification, dynamic system control, prediction and so on for the business process.

In order to investigate the increasing business risks of the enterprise we consider some problems of the identification and the prediction of a discrete model of DFDS and application Genetic Algorithm Technologies in this domain.

The basic approaches to the identification of DFDS that have been developed to this day (2-4,10,11,13,15,16,20-24] and others) can be divided into two groups – analytical and algorithmic – both of which are oriented to a fuzzy process model written in terms of fuzzy compositional or integro-differential equations or their modifications. Various analytical methods and algorithms were used in order to identify such models, i.e., the corresponding relation of spaces of inputs and outputs of fuzzy dynamic systems. These methods mainly imply the construction of some set-theoretic operation that is inverse to the composition operation and requires a subsequent smoothing of the results ([3,4,24] and others). In some works([22,23] and others), fuzzy models of regression type were identified by means of analytical regularization methods that allowed one to obtain numerical estimators of model coefficients. In this paper, a new approach is proposed to the identification of DFDS models, based on the genetic algorithm technology.

Section 2 contains some necessary preliminary concepts of the model definition. Section 3 describes the regularization conditions for obtaining of a quazi-optimal estimator of transition operator of DPDS and a procedure of moving to DPDS finite model. Section 4 describes the use of Genetic Algorithm for solution of identification problem. Obtained results are illustrated by the business process risks' example in Section 5.

II. GENERAL DEFINITION OF THE DFDS

Applying the results of [11,15,16,20], we construct the discrete recurrent fuzzy process with possibility uncertainty, the source of which is a stream of expert knowledge reflections on the states of some evolutionary complex system.

Let's start describing objects of DFDS. Let

$$X = \{x_1, x_2 \dots x_n\}, \quad (1)$$

be the set of possible states of nature of the evolutionary system (business states, situations and so on), observed by expert. The fuzzy trajectory of expert knowledge reflections on the states, with respect fuzzy terms of some linguistic variable (business risks and so on), given in certain time moments, will have the following form:

$$\{\hat{A} = \hat{A}(x, t)\}, t = 0, 1 \dots s, x \in X. \quad (2)$$

Based on the results from [11,15,16,20], let's introduce the following recurrent DFDS of the expert knowledge stream (2):

$$\begin{aligned} \tilde{A}(x, t) &= \int_X \langle \tilde{\rho}(x', x) \bullet \tilde{A}(x', t-1) \rangle \circ \hat{P}os_{t-1}(\cdot); \\ x \in X, t &= 1, 2 \dots s; \end{aligned} \quad (3)$$

$$\tilde{A}(x', 0) \equiv \hat{A}(x', 0), x' \in X.$$

The \int symbol denotes Sugeno's integral [5-10,20] and $\tilde{\rho}$ is some fuzzy transition operator (FTO), which defines transition possibilities [1,11] between states of the system:

$$\tilde{\rho}: X^2 \rightarrow [0,1] \quad (4)$$

In (3) we use possibility measure with the possibility distribution, derived from the expert's knowledge reflections stream (2).

$$\hat{P}os_{t-1}(B) = \frac{\bigvee_{x \in B} \hat{A}(x, t-1)}{\bigvee_{x \in X} \hat{A}(x, t-1)}, \forall B \subset X, t = 1, 2 \dots s. \quad (5)$$

Also, in (3) the symbol \bullet denotes operation, which takes two arguments. For the given $a, b \in [0, 1]$ these operations can be:

1. $a \bullet b = a \wedge b$ (Classical Minimum)
2. $a \bullet b = a \vee b$ (Classical Maximum)
3. $a \bullet b = a \cdot b$ (Algebraic Multiplication)
4. $a \bullet b = a + b - a \cdot b$ (Algebraic Summation)

and so on (See in [2]).

If we successfully identify the transition operator (4) which gives us minimal differences between expert (2) and model (3) data in moments $t = 1, 2 \dots s$, we will be able to predict the system state for the moments of time $t = s + 1, s + 2, \dots$. This can be done using the following formula

$$\begin{aligned} \tilde{A}(x, t) &= \int_X \langle \tilde{\rho}(x', x) \bullet \tilde{A}(x', t-1) \rangle \circ P os_{t-1}(\cdot); \\ x \in X, t &= s + 1, s + 2 \dots \end{aligned} \quad (6)$$

Here we use the possibility measure with the possibility distribution, derived from the model (3). That is:

$$P os_{t-1}(B) = \frac{\bigvee_{x \in B} \tilde{A}(x, t-1)}{\bigvee_{x \in X} \tilde{A}(x, t-1)}, \forall B \subset X, t = s + 1, s + 2, \dots \quad (7)$$

III. FINITE MODEL OF THE DFDS

To restore FTO in the model (1) – (7) we consider some finite possibility recurrent model. In this model, we define the regularization condition for constructing a quasi-optimal estimator of transition operator.

Let's introduce the following finite set of levels:

$$F_N = \left\{ \frac{0}{N}, \frac{1}{N} \dots \frac{N}{N} \right\} \subset [0,1], N \in \mathbb{N}. \quad (8)$$

In model (1)-(7), we consider only those transition operators, which belong to the following set:

$$\Phi_N = \left\{ \tilde{\rho}_N | \tilde{\rho}_N : X^2 \rightarrow F_N \right\}. \quad (9)$$

This way, we have some kind of discretization of transition operator (4). Thus, our goal is to construct the transition operator from Φ_N in which the following deviation achieves its minimum (regularization condition):

$$\Delta = \sum_{x \in X} \sum_{t=0}^s \left| \tilde{A}_{\rho_N}(x, t) - \hat{A}(x, t) \right|. \quad (10)$$

Here we use not the model (3) but its discretization:

$$\begin{aligned} \tilde{A}_N(x, t) &= \int_X \langle \tilde{\rho}_N(x', x) \bullet \tilde{A}_N(x', t-1) \rangle \circ \hat{P}os_{t-1}(\cdot); \\ t &= 1, 2 \dots s; \\ \tilde{A}_N(x', 0) &\equiv \hat{A}_N(x', 0), x' \in X. \end{aligned} \quad (11)$$

It's obvious that even for smaller values of N the cardinality of the set of transition operators (9) is too large. Thus, it is not efficient to simply try all possible variants and take the smallest deviation (10). For this reason, it was decided to solve this problem using a genetic algorithm.

IV. PROBLEM SOLUTION USING GA

Generic algorithms (GAs) operate on a collection of candidate solutions. Each solution is encoded as a finite length string of bits. This string is called chromosome. The set of chromosomes is called population. In classical approach, the initial population is created randomly and GA iterates on it. During the iteration, a fitness function is applied to each chromosome to determine how fit each chromosome is. After selecting appropriate chromosomes, genetic operations (mutation and crossover) take place on them. Finally, the new population is created. After this, GA runs on newly-created population and the process is running until some exit criteria is satisfied. The present paper assumes that reader is familiar with the fundamental concepts of GA.

In our case, we can encode transition operators directly as chromosomes. Obviously, transition operators can be also written as matrices of $n \times n$, containing integers from $[0, N]$ segment. Each of such matrices can be encoded as array of n^2 integer entries. All we need to do now is to convert all these integers to the set of bits to get classical GA

chromosomes. For this, we need to calculate minimal number of bits M^* to store integers from 0 to N there. In our approach, without losing generality, we assume that $N = 2^H - 1$ for some H. Thus, $M^* = H$ in our case. There are 2 advantages of using such numbers. Firstly, we use memory efficiently and secondly, we eliminate risk of getting chromosomes which do not correspond to any valid transition operators. So, we don't have waste computational power on seeking and handling invalid chromosomes.

A fitness function, which is also key part of GA design, is derived directly from the deviation (10) using the following simple transition:

$$FT = \frac{1}{1 + \Delta} \tag{12}$$

We started our experiments with simple roulette-wheel selection. Also, at this time we use fixed-sized populations only.

The model considers frequent use of fuzzy integral, which is calculated several times to finally get chromosome's fitness. As the number of chromosomes within the population is high, it's possible to calculate fitness of all of them in parallel way.

Our solution runs as a single process, on a single multi-core machine. There is a thread pool maintained in this process, responsible to do asynchronous calculation of the fuzzy integrals. It helps us to consume CPU optimally.

The application architecture also allows moving integral calculation to the separate machines. But this feature is still under development.

V. EVALUATION OF THE ENTERPRISE BUSINESS RISKS' DYNAMICS

It is known [7-9,11-13,15,16,19,20] that the models of (1)-(5) types were created to investigate complex, dynamic processes. For example, abnormal and extreme business processes in the uncertainty environment, where it's hard or even impossible to define structures for deterministic or stochastic models. If we perform fuzzification of the investigated system characters, thus introducing new uncertainty, it will be easy for the experts to define structures between system states based on their knowledge and intellectual activities. The knowledge can be reflected in rules, relations, etc. The usage of fuzzy modeling methods is credible, as the method has both heuristic and fundamental basis. We use the same approach. But in the example provided below, the levels of business process states (fuzzy terms of linguistic variable – business risks) are evaluated using fuzzy logic control system. This system is based on the objective statistical data which describes the real process.

We consider the organization (The data is kindly provided by Georgia-based organization, "Fractal LTD") which in dynamics has increasing business risks. To consider the detalization of evolution of the system, we study only this particular characteristic. During the construction of the model, we consider some linguistic variable: "business risks" and its fuzzy terms are considered as the system states. These terms are: x_1 - insignificant business risks; x_2 - low business risks; x_3 -

average business risks; x_4 - significant business risks; x_5 - high business risks.

The fuzzy logic control system was constructed in MatLab Fuzzy ToolBox environment. The system output is defined in fuzzy terms - x_1, x_2, x_3, x_4, x_5 , as fuzzy variables on scale [0-10]. As the system input, we used 15 objective characteristics of the organization - linguistic variables, each defined using 3-, 4-, 5-term fuzzy numbers. The constructed control system was used on each step of modeling (one step corresponded to one month). The results for the previous year, provided as possibility levels, are depicted in Table 1 (The similar problems, but concerning static business risks' evaluations, see in [14,17]).

The monotonic levels in Table 1 show that the risks tend to increase. Our goal is to create model process, which will predict (with high possibility level) an appearance business risks. The result will show, after how many months this risks will appear and what kind of risks it will be. In practice, this is identical to fuzzy stopping problem.

In our case, we have estimations for 12time moments, (Table 1), we use all 12estimations to construct the transition operator and with it try to predict system's states starting from 13-th time moment. So, in model (1)-(11) we have $s = 12$ and $n = 5$.

In modeling, we used Algebraic Summation ($a \bullet b = a + b - a \cdot b$) as \bullet operation.

Table 2 shows modeling results, got after above-mentioned testing. The Fitness value will be $FT=0.7513$, which indicate good correspondence to the observed and modeled values.

As mentioned above, our purpose was to discover the dynamics of business risks of Enterprise with high and increasing possibility. The Table II shows that x_1, x_2, x_4 and x_5 risks don't exhibit increasing dynamics, such dynamics is only shown in x_3 (mean business risks). This means that approximately on the 7-th step (month) of forecasting process the high possibility (0.891 and more) of mean business risks is exhibited.

Thus, in the process of dynamics of business risks of Enterprise, the increasing of possibility of mean business risks is shown, meaning that in future, starting approximately from 7-th month, mean business risks is expected.

TABLE 1. POSSIBILITY LEVELS OF FUZZY TERMS OF THE BUSINESS RISKS

Time moments	Fuzzy logic control system's estimated possibility levels of fuzzy terms of the business risks - $\tilde{A}(x, t)$				
	x_1	x_2	x_3	x_4	x_5
1	0.453	0.351	0.258	0.251	0.204
2	0.461	0.387	0.253	0.264	0.215
3	0.463	0.402	0.366	0.261	0.21
4	0.461	0.437	0.398	0.291	0.209
5	0.484	0.441	0.435	0.319	0.214
6	0.497	0.46	0.493	0.324	0.22
7	0.478	0.491	0.534	0.337	0.233

Time moments	Fuzzy logic control system's estimated possibility levels of fuzzy terms of the business risks - $\tilde{A}(x,t)$				
	x_1	x_2	x_3	x_4	x_5
8	0.48	0.51	0.526	0.374	0.245
9	0.507	0.515	0.681	0.413	0.236
10	0.507	0.55	0.718	0.448	0.249
11	0.495	0.508	0.723	0.404	0.215
12	0.51	0.607	0.743	0.451	0.24

TABLE2. MODELED POSSIBILITY LEVELS OF FUZZY TERMS OF THE BUSINESS RISKS.

Months	Modelled possibility levels of fuzzy terms of business risks - $\tilde{A}(x,t)$				
	x_1	x_2	x_3	x_4	x_5
1	0.423	0.347	0.263	0.291	0.218
2	0.431	0.369	0.312	0.283	0.203
3	0.459	0.471	0.395	0.281	0.217
4	0.473	0.463	0.384	0.293	0.203
5	0.481	0.457	0.423	0.319	0.212
6	0.494	0.463	0.475	0.319	0.21
7	0.494	0.485	0.547	0.327	0.22
8	0.491	0.543	0.626	0.393	0.239
9	0.513	0.549	0.697	0.425	0.241
10	0.508	0.578	0.704	0.437	0.238
11	0.506	0.581	0.723	0.441	0.256
12	0.513	0.582	0.734	0.447	0.256
13	0.508	0.585	0.769	0.448	0.256
14	0.513	0.589	0.726	0.451	0.266
15	0.514	0.591	0.817	0.456	0.271
16	0.521	0.597	0.847	0.459	0.273
17	0.519	0.591	0.823	0.461	0.272
18	0.523	0.606	0.856	0.46	0.279
19	0.529	0.607	0.891	0.463	0.281
20	0.527	0.613	0.916	0.471	0.285
21	0.544	0.619	0.956	0.473	0.287
22	0.556	0.624	0.991	0.474	0.292

At this moment we are working on application of model (1)-(5) in different business problems.

VI. CONCLUSION

The DFDS model can be successfully used in the investigation of monotone business risks' dynamics where the input of the model is the stream of expert knowledge. Because of the recurrent model being non-linear with different compositions, heuristic approach of estimation is best suitable

technique to run the model. As a result the use of Genetic Algorithm proved to be successful tool for running DFDS model.

REFERENCES

- [1] D. Dubois and H. Prade, H. Possibility Theory, Plenum Press, New York, 1988.
- [2] G. J. Klir and M. Higashi, Identification of fuzzy relation systems, IEEE Trans. Systems Man Cybernet. 14, no. 2, 1984, pp. 349-355.
- [3] M. Kurano, M. Yasuda, J. Nakagami and Y. Yoshida, A fuzzy relational equation in dynamic fuzzy systems, Fuzzy Sets and Systems 101, pp. 439-443, 1999.
- [4] D. Ruan, J. Kacprzyk and M. Fedrizzi, *Soft Computing for Risk Evaluation and Management. Applications in Technology, Environment and Finance*, Physica-Verlag, Heidelberg, 2001.
- [5] G. Sirbiladze and A. Sikharulidze, Weighted fuzzy averages in fuzzy environment. I. Insufficient expert data and fuzzy averages, *Internat. J. Uncertain. Fuzziness Knowledge-Based Systems* 11(2), 2003, pp. 139-157.
- [6] G. Sirbiladze and A. Sikharulidze, Weighted fuzzy averages in fuzzy environment. II. Generalized weighted fuzzy expected values in fuzzy environment, *Internat. J. Uncertain. Fuzziness Knowledge-Based Systems* 11(2), 2003, pp. 159-172.
- [7] G. Sirbiladze, Modeling of Extremal Fuzzy Dynamic Systems. Part I: Extended Extremal Fuzzy Measures, *International Journal of General Systems*, 34,2, 2005, pp.107-138.
- [8] G. Sirbiladze, Modeling of Extremal Fuzzy Dynamic Systems. Part II, Extended Extremal Fuzzy Measures on Composition Products of Measurable Spaces. *International Journal of General Systems*, 34,2, 2005, pp.139-167.
- [9] G. Sirbiladze, Modeling of Extremal Fuzzy Dynamic Systems. Part III: Modeling of Extremal and Controllable Extremal Fuzzy Processes, *International Journal of General Systems*, 34,2, 2005, pp.169-198.
- [10] G. Sirbiladze and T. Gachechiladze, Restored fuzzy measures in expert decision-making, *Inform. Sci.* 169(1/2), 2005, pp. 71-95.
- [11] G. Sirbiladze, Modeling of extremal fuzzy dynamic systems. IV. Identification of fuzzy integral models of extremal fuzzy processes, *International Journal of General Systems* 35(4), 2006, pp. 435-459.
- [12] G. Sirbiladze and A. Sikharulidze, Bellman's Optimality Principle in the Weakly Structurable Dynamic Systems, *Advanced Topics on Fuzzy Systems. WSEAS International Conference on fuzzy systems*, Sofia, Bulgaria, 2008.
- [13] G. Sirbiladze, Fuzzy Technologies of Weakly Structurable Systems' Modeling and Simulation, Planary Speech, *Proceedings of the 11th WSEAS International Conference on Automatic Control, Modelling and Simulation (ACMOS '09)*, Istanbul, Turkey, 2009, pp.288-297.
- [14] G. Sirbiladze, I. Khutsisvili, Decision Support's precising Technology in the Investment Project Risk Management, *Proceedings of the 11th WSEAS International Conference on Automatic Control, Modelling and Simulation (ACMOS '09)*, 303-312, Istanbul, Turkey, 2009, pp.303-312.
- [15] G. Sirbiladze, A Fuzzy Identification Problem for the Stationary Discrete Extremal Fuzzy Dynamic System, Planary Speech, *Proceedings of the 3rd International WSEAS Conference on COMPUTATIONAL INTELLIGENCE (CI '09)*, Tbilisi, Georgia, 2009, pp.323-332.
- [16] G. Sirbiladze, Fuzzy Identification Problem for Continuous Extremal Fuzzy Dynamic System, *Fuzzy Optimization and Decision Making*, vol. 9, N. 3, 2010, pp.233-274.
- [17] G. Sirbiladze, I. Khutsisvili and P. Dvalishvili, Decision Precision Fuzzy Technology to Evaluate the Credit Risks of Investment Projects, *IEEE 10-th International Conference on Intelligent Systems Design and Applications (ISDA 2010)*, Cairo, Egypt, 2010.
- [18] G. Sirbiladze and A. Sikharulidze, Generalized Weighted Fuzzy Expected Values in Uncertainty Environment, *Proceeding of the 9-th WSEAS International Conference on Artificial Intelligence, Knowledge*

Engineering and Data Bases, University of Cambridge, UK, 2010, pp.59-64.

- [19] G. Sirbiladze, Fuzzy dynamic programming problem for extremal fuzzy dynamic systems, *In Fuzzy Optimization: Recent Developments and Applications, Studies in Fuzziness and Soft Computing*, W. A. Lodwick and J. Kacprzyk (Eds.), Vol. 254, Phisica-Verlag, Heidelberg, 2010, pp.231-270.
- [20] G. Sirbiladze, Identification problem for continousextremal fuzzy dynamic system, *Fuzzy optimization and decision making*, vol. 9, N. 3, 2010, pp.233-274..
- [21] M. Sugeno, *Theory of fuzzy integrals and its applications*, Ph.D. Thesis of Tokyo Insitute of Technology, 1974.
- [22] H. Tanaka, H. Ishibuchi, and S. Yoshikawa, Exponential possibility regression analysis, *Fuzzy Sets and Systems* 69(3), 1995, pp. 305–318.
- [23] G. Vachkov and T. Fukuda, Simplified fuzzy model based identification of dynamical systems, *Int. J. Fuzzy Systems (Taiwan)* 2(4), 2000, pp. 229–235.
- [24] Y. Yoshida, Markov chains with a transition possibility measure and fuzzy dynamic programming, *Fuzzy Sets and Systems* 66, 1994, pp. 39–57.

An Empirical Study of Tracking Strategies of E-commerce Websites

Wasin Treesinthuros

Graduate School of Computer and Engineering Management
Assumption University of Thailand
Bangkok, Thailand
P5319426@au.edu

Abstract— This research looks into the factors affecting the tracking strategies of e-commerce websites. Each of these factors is analyzed into a deeper view with every factor being sought to reveal its relevance to the topic of study. The factors that have been taken into consideration as having affected the tracking strategies of e-commerce websites include: conversion rate, cost per visitor, order conversion rate, buyer conversion rate, product conversion rate and shopping cart abandonment rate. The research further looks at what other authors have researched and concluded about the factors examined as having an effect on tracking strategies of e-commerce websites. The research elaborates how these factors have influenced website tracking strategies according to different authors, researchers and corporations. Data analyzed in the research was collected using the Likert scale and was analyzed using the AMOS analysis method. All measures were based on items in existing instruments and literature. The researcher used 5 point Likert-Scales where 1 represents “strongly disagree” and 5, “strongly agree.” The method is confirmed empirically using confirmatory factor analysis. The research ends with a conclusive summary of the finding of the research on the empirical study of tracking strategies of e-commerce websites.

Keywords- e-commerce, internet marketing, e-business, e-marketing, website tracking, tracking strategy

I. INTRODUCTION

Site tracking, with reference to e-commerce, refers to the process of closely analyzing all the activities being carried out by the user of an e-commerce marketing website, with an aim to realize the success or failure of the marketing strategies adopted by an e-commerce marketing website.

In e-commerce, website marketing is the backbone of the success of business on the electronic platform, since technology has now made it possible that a seller and a buyer in the market do not have to have physical contact to conduct business. It is essential for an e-commerce website to have a track on the sales leads, the website visibility and all other factors that lead to the success of e-commerce marketing. Having a website with loads of traffic is just a step into the real world of e-commerce, but having a strategy to convert the traffic into leads is the core value of marketing on the e-commerce platform.

It is necessary to have your e-commerce website listed and ranked highly by search engines; however, the resultant high traffic does not guarantee a sale. Thus, it is not the ultimate marketing strategy to employ in the e-business.

It is said that there is a word of difference between your advertisement being delivered to the client and your advertisement being noticed [8]. This research, therefore, looks into the factors that affect the tracking strategies of an e-commerce website and how they can be adjusted to meet the marketing team’s goals.

II. LITERATURE REVIEW

Tracking strategies should aim at converting traffic from an e-commerce into a lead sale and thus making a sale or potential client from it. This should be the core measure of e-commerce marketing success. A lead may be defined as the pre-qualified visitor emerging from a marketing campaign. Information on a lead is provided by web analytics tracking an e-commerce marketing campaign [7].

Tracking a website user’s activity helps the website owner to know which areas to improve on the website. It is through website tracking that the website owner can know which page contents seem to be more appealing and more relevant to the users. The tracking strategies of e-commerce websites that have been analyzed in this research paper include:

A. Conversion rate

This may be defined as the percentage of your site’s traffic that places an order, sends an inquiry, subscribes to your newsletter, calls you or comes by to visit your business. It may also be referred to as the percentage of website visitors that transform into lead sales by providing their contact details by filling out a web form [8]. Conversion rate is affected by the nature of your customer and your offering. A good conversion rate, according to Michael [8], is from 5 to 25 percent. Conversion rate should be constantly improved by developing a more user friendly site and content, analyzing keenly the website’s traffic sources and reducing the average bounce rates [6].

The success of a conversion rate improvement strategy may be tested using the split test method, whereby a control sample is tested against single tests to see the resultant conversion rate and determine the improvement. In addition, a more complex method may be applied; multivariate testing. This is the kind of testing that tests multiple components of a website in a live environment [10]. The ultimate goal of a conversion marketer should be to increase traffic to the website and then increase the conversion rate of the visitors into buyers.

B. Cost per visitors

Brian [1] used a mathematical example to explain the relationship between the cost per visitor and the conversion rate of a website. A first time visitor to a website can on average cost between \$.30 and \$20.0 per visit time. On the assumption that a first time visitor costs \$2.00 per visit with the consideration of time, promotional and advertising costs, overhead costs and design work, the conversion rate is at 2 percent. Each newly acquired customer each costing \$100 per transaction, even in a case where the customer spends \$29.97 per transaction.

The high cost per visitor results to a lower conversion rate. To increase the conversion rate, the marketer needs to lower the cost per visitor. Increase in conversion rate is directly proportional to the increase in the number of items sold.

C. Order conversion rate

This is a website performance tracking strategy that employs the use of websites using a unique script added to the order confirmation page to track the number of orders, quantity and the amount [9]. According to Goldwyn [4], most online firms are able to convert only 2 to 3 percent of their traffic to orders. Increasing the order conversion rate should be the core business of any tracking strategy for all e-commerce websites.

D. Buyer conversion rate

The main aim of tracking the buyer conversion rate is, to monitor the movement of the site visitor from the origin point to the point of entry into the website, to their navigation through the site, drop off and the possible return [2]. According to Purtsis and Narasimhan [12], this strategy is based on the buying behavior of a consumer and can be analyzed basing on their cultural, social, and economic backgrounds. Buyers make decisions to buy having considered several factors. In buyer conversion rate, these factors should be tracked to evaluate the buyer response.

E. Product page conversion rate

The main problem affecting the product page conversion rate is the rate at which potential customers abandon the product page without actually checking out to the payment page. One of the factors that may be affecting the rate at which pages are abandoned is the page response time. Most online customers wish to save time that is why they do not go to the stores physically. Therefore, a shopping page that is slow in loading products will be a waste of time leading to customers abandoning the page. Increase in the page loading speed will directly result in a decrease in the rate of product page abandonment. Figure 1 below shows the comparison between rates of page abandonment against the change in speeds of page responses [5].

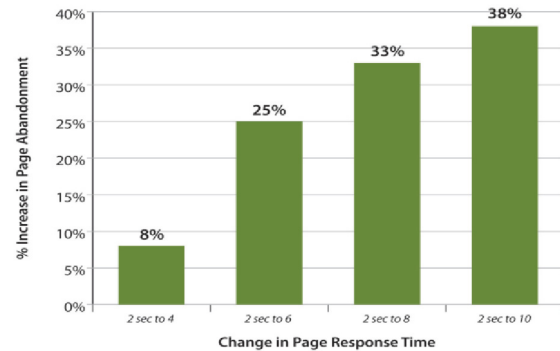


Figure 1. Change in page abandonment against change in page response time

F. Shopping cart abandonment rate

The shopping cart abandonment rate refers to the percentage of buyers who go through the first stages of online buying but do not end up at the payment point. There are several factors that can lead to this kind of behavior by online buyers. One of them is by enhancing the ease of navigation on the shopping page. The other factor is by comparison shopping, whereby the customer compares prices with other online shops. The web designer should ensure that the only easiest way to exit the shopping page should be by checkout to the payment page [3]. This should also be simple to do as shown in figure 2 below;

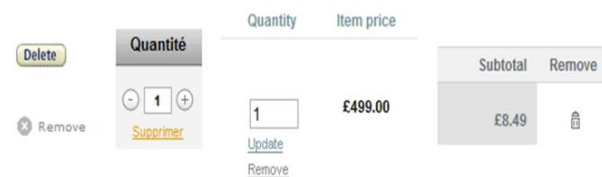


Figure 2. Simple and user friendly online shopping process

The use of a promotional strategy to avoid shopping cart abandonment to keep the client glued to the free gift received after purchase may be employed to curb the consumer behavior of abandoning the shopping cart. However, according to a research by Oliver and Shor [11], the role of online promotional coupons may serve as a factor resulting in shopper dissatisfaction and the result is shopping cart abandonment. They found out that customers were dissatisfied with the request to enter the promotional code which in most cases they did not have, thus ended up abandoning the shopping cart.

III. EMPIRICAL RESULTS

All measures were based on items in existing instruments and literature. The researcher was measured using 5 point Likert scales where 1 represents "strongly disagree" and 5, "strongly agree." Through the internet, we collected 400 valid records.

A. Confirmatory Factor Analysis

The researcher has created the conceptual framework from the literature into figure 3. This model has one latent variable “Site tracking” (SITTRK) and six observe variables; “Conversion Rate” (CONR), “Cost per visitor” (CPV), “Order conversion rate” (OCR), “Buyer conversion rate” (BCR), “Product page conversion rate” (PPCR) and “Shopping cart abandonment rate” (SCAR).

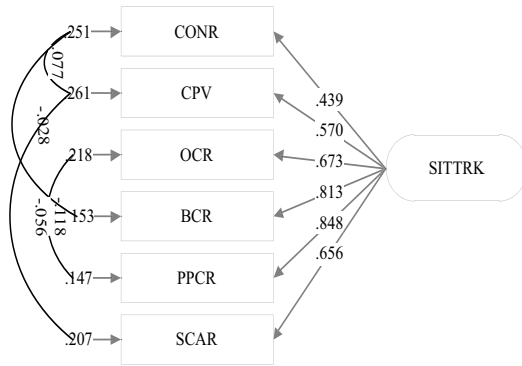


Figure 3. Site tracking factors for e-commerce model (Output from AMOS graphic)

The researcher selected CFA (Confirmatory Factor Analysis) as the basis to confirm the relationship between “Site tracking” (SITTRK), and “Conversion Rate” (CONR), “Cost per visitor” (CPV), “Order conversion rate” (OCR), “Buyer conversion rate” (BCR), “Product page conversion rate” (PPCR) and “Shopping cart abandonment rate” (SCAR). The numbers represent the regression weight which indicates the degree of relationship between variables. The relationship between product page conversion rate and site tracking is the highest at 0.848. This means that the product page conversion rate is the most affected and least productive. In the other words, the site owner has to focus on the product page conversion rate and find the appropriate strategies to improve the product page and in turn the entire website’s performance.

This model does not only show the degree of the relationship between variables but also calculates the model fit summary for implementing into the business.

B. Model fit summary

The equation (1) represents that “Product conversion rate” is more effective than “Conversion rate”, “Cost per visitor”, “Order conversion rate”, “Buyer conversion rate” and “Shopping cart abandonment rate”.

$$F(\text{SITTRK}) = 0.439f\text{CONR} + 0.570f\text{CPV} + 0.673f\text{OCR} + 0.813f\text{BCR} + 0.848f\text{PPCR} + 0.656f\text{SCAR} \quad (1)$$

TABLE I. CMIN

Model	NPAR	CMIN	DF	P	CMIN/DF
Default model	16	8.823	5	.116	1.765

In this model, the number of parameters is 16, degrees of freedom are 5, the probability of getting a discrepancy as large as 8.823 are 0.116 and relative chi-square (CMIN/DF) is 1.765. Therefore, the result represents a perfect fit.

TABLE II. PARSIMONY-ADJUST MEASURES

Model	PRATIO	PNFI	PCFI
Default model	.333	.330	.332

Parsimony ratio (PRATIO) of this model is 0.333 that indicates a very good fit.

TABLE III. NCP

Model	NCP	LO 90	HI 90
Default model	3.823	.000	16.245

The columns labeled LO 90 and HI 90 contain the lower limit (δ_l) and upper limit (δ_u) of 90% confidence interval for δ . With approximately 90 percent confidence, the population NCP for this model is 3.823.

TABLE IV. BASELINE COMPARISON

Model	NFI	RFI	IFI	TLI	CFI
Default model	.990	.970	.996	.987	.996

NFI, RFI, IFI, TLI and CFI values close to 1 indicate a very good fit. For this model, Normed Fit Index (NFI) is 0.990, Relative Fit Index (RFI) is 0.970, Incremental Fit Index (IFI) is 0.996, Tucker-Lewis coefficient Index (TLI) is 0.987 and Comparative Fit Index (CFI) is 0.996. Therefore, the result from table IV indicates the model has a very good fit.

TABLE V. FMIN

Model	FMIN	F0	LO 90	HI 90
Default model	.022	.010	.000	.041

With approximately 90 percent confidence, the population F0 for this model is 0.022, between 0.000 and 0.025.

TABLE VI. RMSEA

Model	RMSEA	LO 90	HI 90	PCLOSE
Default model	.044	.000	.090	.520

RMSEA (root mean square error of approximation) of 0.080 or less indicates a close fit. With approximately 90 percent confidence, the population RMSEA for this model is 0.044. Therefore, the result from table VI indicates the model has a very close fit.

IV. CONCLUSION

It is vital for e-marketers to monitor and evaluate the response of their marketing strategies in the market. It is not

important to have a website which is not visible to customers, at the same time; a website with so much traffic that does not get converted to sales is also of no use. Advertisement is supposed to catch the customers' attention, and then the attention is supposed to drive the customer into making a purchase. It is therefore necessary to implement software and all supportive machinery that will ensure that the activity of any visitor on the e-commerce website is monitored and analyzed. This way, the markets will know which areas to improve to meet the clients' need as marketing is all about providing products that fulfill the clients' needs.

The findings of this research show the weight of effect between the tracking strategies. The e-commerce department has to consider that the importance of tracking strategies depends on the weight of effect on the following; "Product page conversion rate", "Buyer conversion rate", "Order conversion rate", "Shopping cart abandonment rate", "Cost per visitor" and "Conversion rate". Most e-marketers only consider the "Conversion rate" because this number can track the overall performance of an e-commerce site but when the e-marketer has to carefully consider the revenue, there are more specific tracking strategies that can improve the site performance and decrease the time used in site improvement.

The e-commerce website tracking strategies that are employed should aim to identify the loopholes into having traffic that does not reflect on the sales side of the equation. This way, improved consumer products will be availed and the e-commerce website will be a success.

REFERENCES

- [1] H. Brian. (n.d.). *Conversion marketing: The psychology of converting browsers into buyers*. [Online]. Available: <http://www.conversionmarketingbook.com/ConversionMktgChapter-1.pdf>
- [2] Clickstream Technologies. (n.d.). *Digital multichannel marketing- the case for improved tagging systems; why dat accuracy matters, where inaccuracies comes from, how they impact in online marketing and how they affect your business*. [Online]. Available: <http://www.clickstream.com/t3/downloads/pdf/P2.pdf>
- [3] Author Unkown. (n.d.). *Analyzing your traffic; discover your sites findability triumphs and tragedies with traffic analysis systems*. [Online]. Available: http://buildingfindablewebsites.com/d/chapters/13_BFW.pdf
- [4] Conversion medic. (n.d.). *Ten ways to improve e-commerce shopping cart conversion rate*. [Online]. Available: <http://www.conversionmedic.com/basket-optimization/>
- [5] C. Goldwyn. (2003). *The art of the cart: Vividence corporation report*. [Online]. Available: <http://www.docstoc.com/docs/20153981/Shopping-Cart-Abandonment-at-Retail-Websites---A-Multi-Stage-Model-of>
- [6] Gomez. (2010). *Why web performance matters- Is your site customer people away?* [Online]. Available: http://www.gomez.com/pdfs/wp_why_web_performance_matters.pdf
- [7] H. Jerry. (2005). *Website conversion rates, more strategies*. Hart Creative Marketing Inc. Concord, California. [Online]. Available: <http://www.hartcreativemarketing.com/marketing-articles/website-conversion-strategy.pdf>
- [8] S. Marshal. (2004). *Search engine effectiveness metrics and score carding*. [Online]. Available: <http://www.searchengineworkshops.com/SearchEngineMetrics.pdf>
- [9] O. Michael. (n.d.). *Click to lead; the website challenge; how to reach business software buyers and convert them to sales ready leads*. [Online]. Available: <http://www.capterra.com/docs/capterra-website-challenge.pdf>
- [10] Net Applications. (n.d.). *Tracking your campaigns with hitlinks enterprise*. [Online]. Available: <http://www.hitslink.com/whitepapers/maximize-your-ROI.pdf>
- [11] Net Applications. (1999). *Content optimization with hits link: Improve conversion rate through multivariate and A/B testing*. [Online]. Available: <http://www.hitslink.com/whitepapers/Content-optimization.pdf>
- [12] R. L. Oliver & M. Shor, "Digital redemption of company satisfying and dissatisfying effects of promotional codes," *J. Product and Brand Management*, Vol. 12, pp. 121-134.
- [13] W. P. Purtsis & S. Narasimhan. (1994). *Analyzing consumer markets and buyer behavior*. [Online]. Available: <http://www.scribd.com/doc/19074691/Buyer-Bhavior-1>

Automation of Business-Processes of an Election System

Gia Surguladze

Dept. of Management Information Systems
of Georgian Technical University
Tbilisi, Georgia
g.surguladze@gtu.ge

Nino Topuria

Dept. of Management Information Systems
of Georgian Technical University
Tbilisi, Georgia
nino.topuria@gtu.ge

Ekaterine Turkia

Dept. of Management Information Systems
of Georgian Technical University
Tbilisi, Georgia
e.turkia@gtu.ge

George Basiladze

Dept. of Management Information Systems
of Georgian Technical University
Tbilisi, Georgia
gbasiladze@yahoo.com

Abstract— In the thesis we consider a critical analysis of foreign and Georgian existing election systems and is described the new, conception of development of support IT infrastructure of an Electronic Election System, proposed by us. We process the bpmn diagrams of business processes of a traditional and electronic election system. Is analyzed their strengths and weaknesses. We offer the conceptual model of databases of an electronic election system, projected in ORM instrument. Is worked out IT supported client-server and service oriented architecture for a system.

Keywords- *Electronic Election; System;Automation;Business-Process; BPMN; ORM; ERM; DDL.*

I. INTRODUCTION

The processes of building of democratic State and a formation of legal society are various depending on many objective and subjective factors. One of the most important from them is the reform of a State Election System. It is especially for our Country with a broad political spectrum and a strong characteristic of non confidence population. We find actual the political, economical and technical-technological decisions to solve the problems of this sphere. In the report we consider a critical analysis of foreign and Georgian existing election systems and is described the new, conception of development of support IT infrastructure of an Electronic Election System, proposed by us [1].

Since 1990, many elections were conducted in Georgia, as local governance, parliament and a presidential also. Referendums were carrying on as well. But the population in spite of many elections is very non confidence of all of them. It was impossible to improve a Georgian legislation and a system of election. There is nothing to proclaim regarding a voters list and the formation of it, where we can find a lot of death people's names. Besides, to conduct an election with an existing election system, costs a lot of money for state.

II. BUSINESS PROCESSES OF ELECTION SYSTEM

We worked out an electronic election system, which excludes all inaccuracies and all misunderstandings, comparing with an old election system. It conditions high quality transparency and develops the confidence in a population. There are several components to be done for the system, which we propose. Each component includes some subparagraphs:

1. Develop a network on base of statement security standards, which will cover all districts, where the elections were founded with an old election system;
2. Equip all electoral districts with a hardware, which will be set in a above described network;
3. Create data warehouses in parallel with a public registry, which databases will be compared and combined with a public registry databases;
4. Major changes will be held on an electoral district. For example: in our conception described election system excludes marking system, ballot-paper and electoral bin too. These processes are described by BPMN models [2]: traditional (1) and electronic election system (Fig.1,2)
5. Develop the software system, which will provide the identification except of existing text databases, multimedia databases also, like a bio photo and fingerprints.

Traditional system foresees following procedures: 1. to find out Name and Surname in an election list; 2. the detection for markers; 3. registration; 4. getting of voting-paper; 5. voting for a preferred candidate; 6. placing the voting paper in a ballot-box. 7. Leaving the voting place.

Electronic election system foresees following procedures: 1. Registration to any registrant with a following method: after that the person is registered and the text data is inputted in a database, they should make a finger print and taking the bio

photo. The finger prints and bio photos should be entered in a database too. 2. After the successful identification the citizen will get an access to go to activated monitor with a list of candidates in a ballot cabinet; 3. Voting for a preferred candidate; Man should point on a touch screen with a special pen; 4. Leaving the voting place;

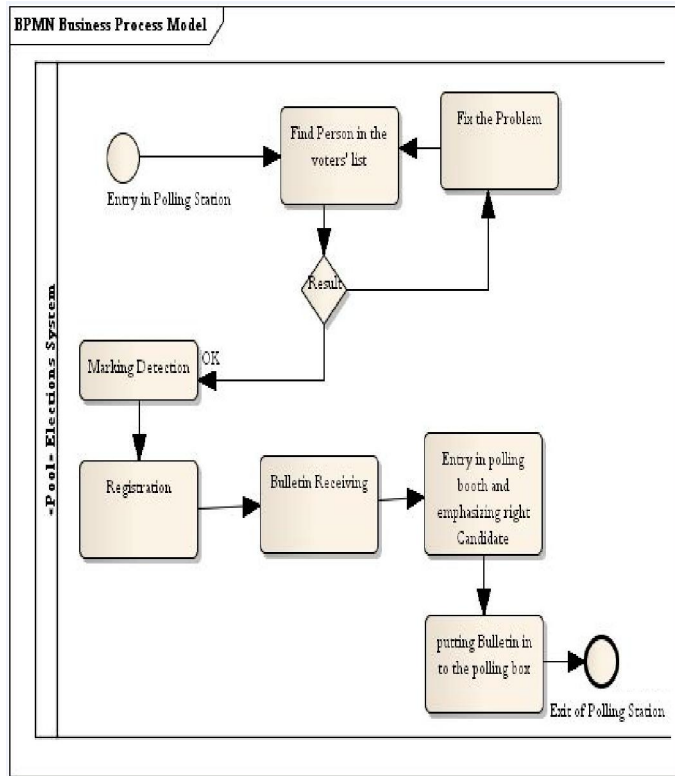


Figure 1. Traditional election system

III. BUILDING OF CONCEPTUAL MODEL –ORM/ERM

Object-Role Modeling (ORM) is primarily a method for modeling and querying an information system at the conceptual level. ORM is so-called because it pictures the world in terms of objects (entities or values) that play roles (parts in relationships).

Unlike Entity-Relationship (ER) modeling and Unified Modeling Language (UML) class diagrams, ORM treats all facts as relationships (unary, binary, ternary etc.).

NORMA (Natural ORM Architect for Visual Studio) is a free and open source plug-in to Microsoft Visual Studio. It supports ORM 2 notation, and can be used to map object-role models to a variety of implementation targets, including major database engines, object-oriented code, and XML schema. [3]

First of all we analyze problematic demands area, definition of a technical assignment. This is the formation of facts. From these elementary facts are defined the ORM-model. Afterwards we build the ORM-diagram.

Initially we analyze the requirements of a technical task of the problem area, from where established the facts. The basic facts are determined by means of the ORM-model, which will be build after the ORM-Diagram (Fig.3).

Elementary facts:

- F1 Voter *has* VoterName
- F2 Voter *has* FingerPrint
- F3 Voter *voted* for a Majoritary_Deputy
- F4 Majoritary_Deputy *has* Deputy_Name
- F5 Voter *votes* in a Polling_District
- And etc.

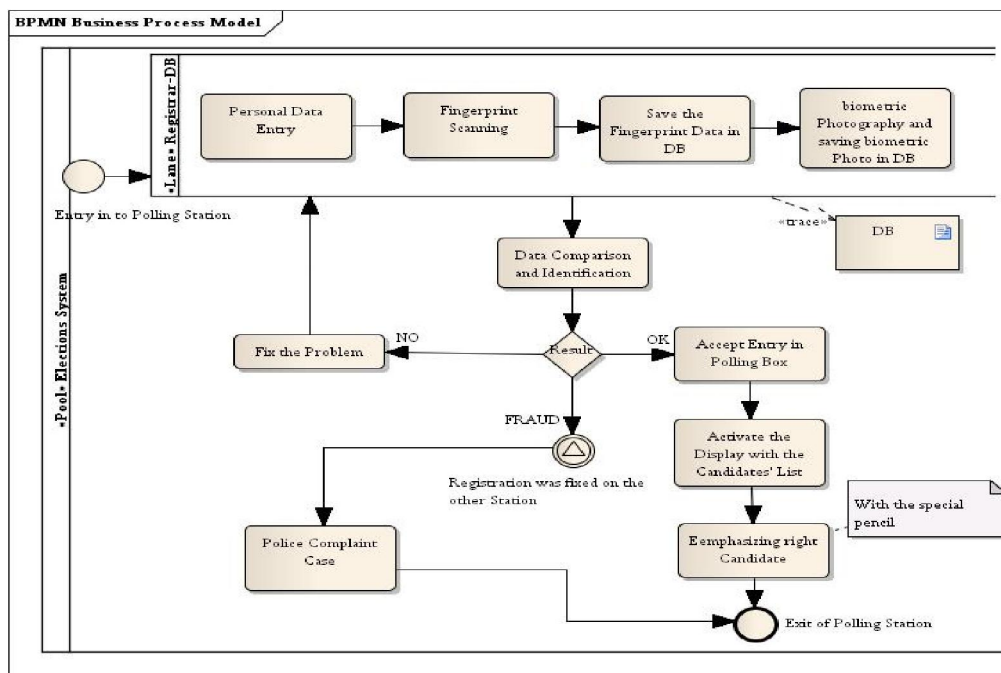


Figure 2. Electronic election system

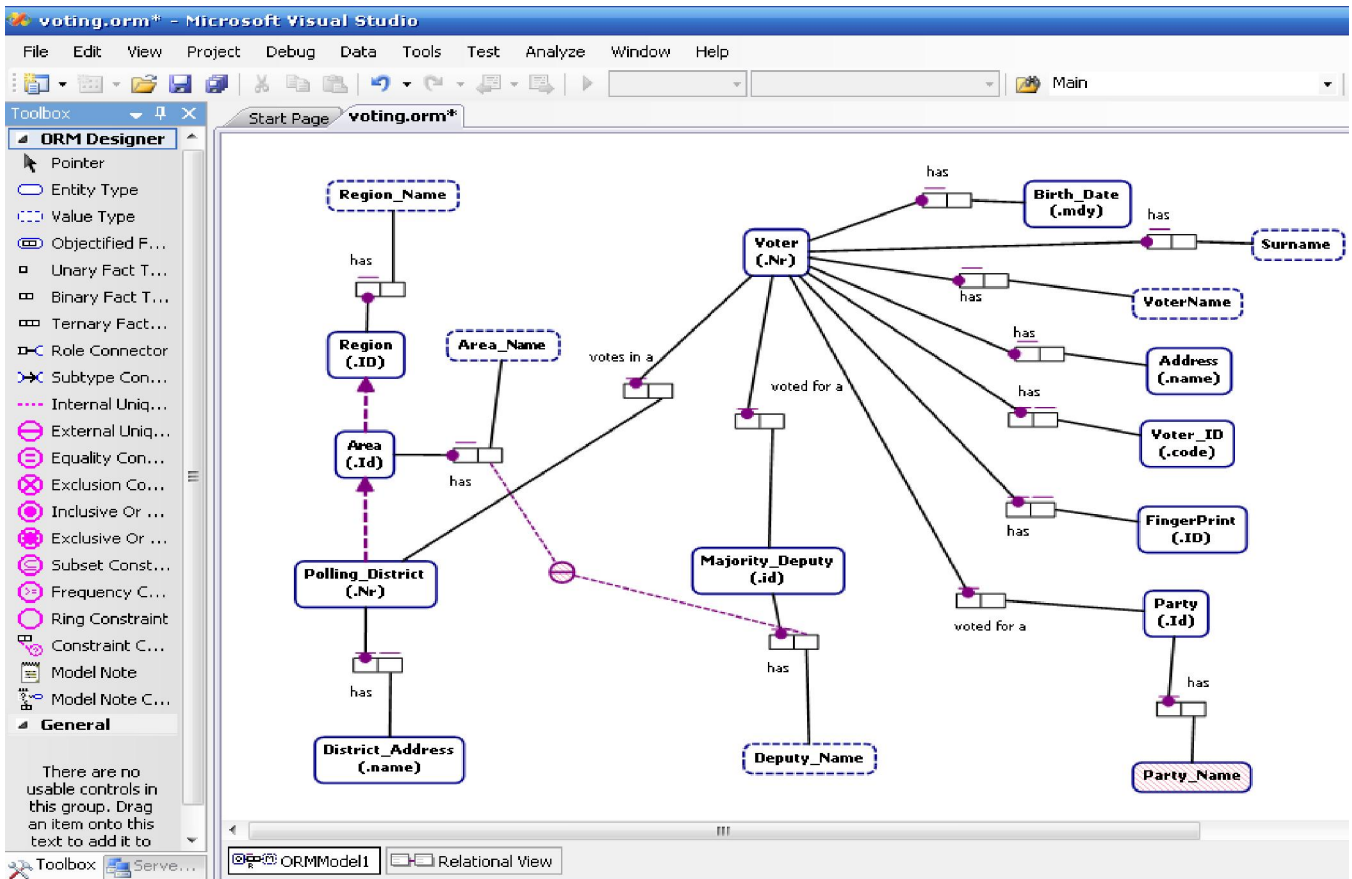


Figure 3. ORM with Objects and Predicates

Once the conceptual schema has been specified Norma tool promote obtain Entity Relation (ER) Model. As seen before, the ORM schema maps to a 3 table relational schema (Fig.4). Predicates describe here the following facts:

Voter is an entity type.

Reference Scheme: Voter has Voter_Nr.

Reference Mode: .Nr.

Fact Types:

Voter has Voter_Nr.

Voter has VoterName.

Voter has Surname.

Voter has Address.

Voter has FingerPrint.

Voter voted for a Party.

Voter voted for a Majority_Deputy.

Voter has Voter_ID.

Voter has Birth_Date.

Voter votes in a Polling_District.

VoterName is a value type.

Portable data type: Text: Variable Length.

Fact Types:

Voter has VoterName.

Voter has VoterName.

Each Voter has exactly one VoterName.

It is possible that more than one Voter has the same VoterName.

Voter has Surname.

Each Voter has exactly one Surname.

It is possible that more than one Voter has the same Surname.

Address is an entity type.

Reference Scheme: Address has Address_name.

Reference Mode: .name.

Fact Types:

Address has Address_name.

Voter has Address.

Voter has Address.

Each Voter has exactly one Address.

It is possible that more than one Voter has the same Address.

Surname is a value type.

Portable data type: Text: Variable Length.

Fact Types:

Voter has Surname.

FingerPrint is an entity type.

Reference Scheme: FingerPrint has FingerPrint_ID.

Reference Mode: .ID.

Fact Types:

FingerPrint has FingerPrint_ID.

Voter has FingerPrint.

Voter has FingerPrint.

Each Voter has exactly one FingerPrint.

For each FingerPrint, at most one Voter has that FingerPrint.

Voter voted for a Party.

Each Voter voted for a exactly one Party.

It is possible that more than one Voter voted for a the same Party.

Party is an entity type.

Reference Scheme: Party has Party_Id.

Reference Mode: .Id.

Fact Types:

Party has Party_Id.
 Voter voted for a Party.
 Party has Party_Name.
 Party has Party_Name.
Each Party has **exactly one** Party_Name.
It is possible that more than one Party has **the same** Party_Name.
 Majority_Deputy is an entity type.
Reference Scheme: Majority_Deputy has Majority_Deputy_id.
Reference Mode: .id.

Fact Types:
 Majority_Deputy has Majority_Deputy_id.
 Voter voted for a Majority_Deputy.
 Majority_Deputy has Deputy_Name.
 Voter voted for a Majority_Deputy.
Each Voter voted for a **exactly one** Majority_Deputy.
It is possible that more than one Voter voted for a **the same** Majority_Deputy.
 Deputy_Name is a value type.
Portable data type: Text: Variable Length.

Fact Types:
 Majority_Deputy has Deputy_Name.
 Majority_Deputy has Deputy_Name.
Each Majority_Deputy has **exactly one** Deputy_Name.
It is possible that more than one Majority_Deputy has **the same** Deputy_Name.
 Voter_ID is an entity type.
Reference Scheme: Voter_ID has Voter_ID_code.
Reference Mode: .code.

Fact Types:
 Voter_ID has Voter_ID_code.
 Voter has Voter_ID.
 Voter has Voter_ID.
Each Voter has **exactly one** Voter_ID.
For each Voter_ID, **at most one** Voter has **that** Voter_ID.
 Birth_Date is an entity type.
Reference Scheme: Birth_Date has Birth_Date_mdy.
Reference Mode: .mdy.

Fact Types:
 Birth_Date has Birth_Date_mdy.
 Voter has Birth_Date.
 Voter has Birth_Date.
Each Voter has **exactly one** Birth_Date.
It is possible that more than one Voter has **the same** Birth_Date.
 Region is an entity type.
Reference Scheme: Region has Region_ID.
Reference Mode: .ID.

Fact Types:
 Region has Region_ID.
 Region has Region_Name.
Each Area is an instance of Region.
 Region_Name is a value type.
Portable data type: Raw Data: Variable Length.

Fact Types:
 Region has Region_Name.
 Region has Region_Name.
Each Region has **exactly one** Region_Name.
It is possible that more than one Region has **the same** Region_Name.

Area is an entity type.
Reference Scheme: Area has Area_Id.
Reference Mode: .Id.

Fact Types:
 Area has Area_Id.
 Area has Area_Name.
Each Polling_District is an instance of Area.
Each Area is an instance of Region.
 Area_Name is a value type.
Portable data type: Text: Variable Length.

Fact Types:
 Area has Area_Name.
 Area has Area_Name.
Each Area has **exactly one** Area_Name.
It is possible that more than one Area has **the same** Area_Name.
 Polling_District is an entity type.
Reference Scheme: Polling_District has Polling_District_Nr.
Reference Mode: .Nr.

Fact Types:
 Polling_District has District_Address.
Each Polling_District is an instance of Area.
 Polling_District has Polling_District_Nr.
 Voter votes in a Polling_District.
 District_Address is an entity type.
Reference Scheme: District_Address has District_Address_name.
Reference Mode: .name.

Fact Types:
 District_Address has District_Address_name.
 Polling_District has District_Address.
 Polling_District has District_Address.
Each Polling_District has **exactly one** District_Address.
For each District_Address, **at most one** Polling_District has **that** District_Address.
 Voter votes in a Polling_District.
Each Voter votes in a **exactly one** Polling_District.
It is possible that more than one Voter votes in a **the same** Polling_District.
Context: Area has Area_Name; Majority_Deputy has Deputy_Name.

In this context, each Area_Name, Deputy_Name **combination is unique.**

Each Area is an instance of Region.
Each Polling_District is an instance of Area.

IV. BUILDING OF DATABASE DDL-FILE

NORMA software generates the DDL code to create the relational schema. Solution Explorer show a code currently generated for SQL Server:

```
CREATE VIEW ORMModel1.Region_UC1 (areaId)
WITH SCHEMABINDING
AS
    SELECT areaId
    FROM ORMModel1.Region
    WHERE areaId IS NOT NULL
GO
```

```

CREATE UNIQUE CLUSTERED INDEX Region_UC1Index ON
ORMModel1.Region_UC1(areald)
GO
CREATE VIEW ORMModel1.Region_UC2(
polling_DistrictDistrict_AddressName
WITH SCHEMABINDING
AS
SELECT polling_DistrictDistrict_AddressName
FROM ORMModel1.Region
WHERE polling_DistrictDistrict_AddressName IS NOT NULL
GO
CREATE TABLE ORMModel1.Voter
(
voterNr INTEGER NOT NULL,
fingerPrintID INTEGER IDENTITY (1, 1) NOT NULL,
Voter_IDCode NATIONAL CHARACTER(4000) NOT NULL,
voterName NATIONAL CHARACTER
VARYING(MAX) NOT NULL,
surname NATIONAL CHARACTER VARYING(MAX) NOT
NULL,
addressName NATIONAL CHARACTER VARYING(MAX)
NOT NULL,
partyId INTEGER IDENTITY (1, 1) NOT NULL,
birth_DateMdy NATIONAL CHARACTER VARYING(MAX)
NOT NULL,
majority_DeputyId INTEGER NOT NULL,
polling_DistrictNr INTEGER NOT NULL,
CONSTRAINT Voter_PK PRIMARY KEY(voterNr),
CONSTRAINT Voter_UC1 UNIQUE(fingerPrintID),
CONSTRAINT Voter_UC2 UNIQUE(Voter_IDCode)
)
GO
CREATE TABLE ORMModel1.Majority_Deputy
(
majority_DeputyId INTEGER IDENTITY (1, 1) NOT NULL,
deputy_Name NATIONAL CHARACTER VARYING(MAX)
NOT NULL,
CONSTRAINT Majority_Deputy_PK PRIMARY
KEY(majority_DeputyId)
)
GO

```

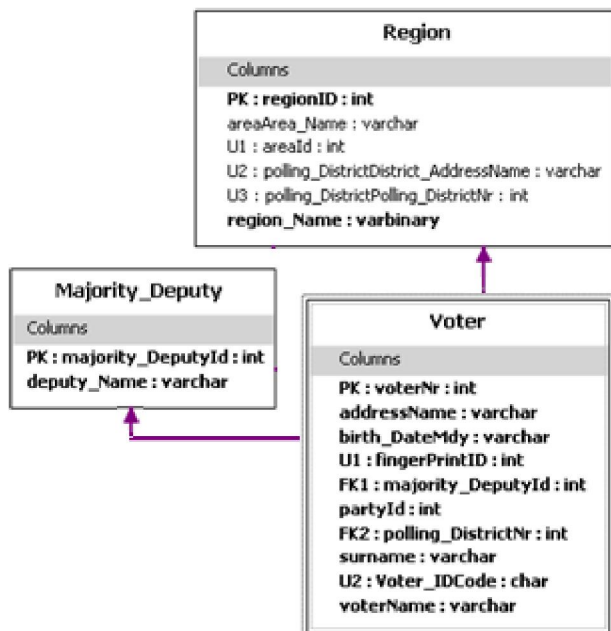


Fig.4. ERM

V. PROCEDURE OF DATABASE BUILDING

NORMA software generates the generates the DDL code to create the relational schema. Solution Explorer show a code currently generated for SQL Server (Fig.5).

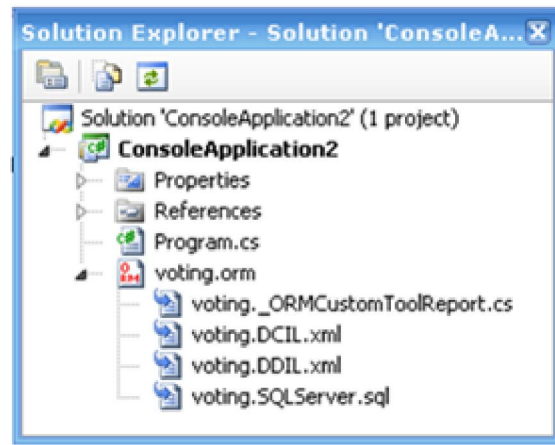


Figure 5. DB Ms SQL Server

VI. CONCLUSION

However, the ORM instrument describes the system on the data structure level only and does not provide the determination of the process structures and scenario-relevant system behavior. To describe the workflow and business process lifestyle, the BPMN model is applied. Described approach of the business modeling is as useful for Change Management as realization of complex Software systems and integration of various applications processes.

We can negotiate about economic efficiency of the conception. If we will foresee, that the basic costs should be paid only once for getting a hardware and creating a software and after that we will need only the installation and dismantle fees, we can say, that it will gone a be cheaper for the state comparing with an old system. State does not need to print ballot-papers, which will be colossal benefit for the state budget. The social effect should be marked. The system will make population more confidence to the election. It gives long term guarantee to the citizen, going to the election implementing his constitutional right with a hope, that his vote will not be lost.

REFERENCES:

- [1] G. Basiladze. Development of Support IT-infrastructure of an Electronic Election System. Transact. of Georgian Technical University. Automated Control Systems - No 1(8), Tb., 2010. pp.223-226.
- [2] E. Turkia, Automation of Business-Project Management Technological Process. Publishing house "Technical University", ISBN 978-9941-14-784-5. Tb., 2010.
- [3] T. Halpin, ORM-2 Graphical Notation. Neumont University, 2005. http://www.orm.net/pdf/ORM2_TechReport1.pdf

On the new multistage Fuzzy Technology to Investment Decisions

Gia Sirbiladze

Department of Computer Sciences
Iv.Javakhishvili Tbilisi State University
Tbilisi, Georgia
gia.sirbiladze@tsu.ge

Irina Khutsishvili

Department of Computer Sciences
Iv.Javakhishvili Tbilisi State University
Tbilisi, Georgia
irina.khutsishvili@tsu.ge

Bezhan Ghvaberidze

Department of Computer Sciences
Iv.Javakhishvili Tbilisi State University
Tbilisi, GEORGIA
bezhan.ghvaberidze@tsu.ge

Abstract— This article proposes a novel Fuzzy Technology to support the Investment Decisions. While choosing among competitive investment projects, this technology provides the selection of projects with minimal crediting risks, makes ranking of chosen projects and then allows to optimally allocate investment amounts between several of them. The technology combines two fuzzy-statistical methods and solution of bicriteria discrete optimization problem, providing three stages of investment projects' evaluation.

Keywords - Investment project risks; expert estimates; positive and negative measures of discrimination; experton; possibility distribution; bicriteria discrete optimization problem.

I. INTRODUCTION

A financial activity of banks, investment funds comes with the risk of the loss, especially in crediting. Thus, the issue of increasing the effectiveness of credit policies and lowering credit risks becomes principal [11,20 and others].

The investment decision-making usually uses special methods. The further development in the field was the probabilistic approach to the evaluation of risks of investment decisions. The methods also can be based on the possibility analysis [2].

Along with traditional statistical techniques, new credit scoring models are developed to support credit decisions. The investment decision-making is influenced by the various uncertainty factors and, the need to formalize and process fuzzy, insufficient and, mainly, expert data. Ignoring the above mentioned factors results in inadequate and non-acceptable decisions. Correct processing of such data is provided with application of fuzzy-set approach [1,7, 9-21,23,24 and others].

Literature, published for the past decade, proposes application of fuzzy-statistical models, neural and fuzzy-neural networks and genetic algorithms when evaluating credit risks [8,20,25 and others]. All of the approaches mentioned above are based on the objective databases and expert data.

The authors are experienced in applying heuristic methods to the decision-making problems which are based on the objective and expert data [8,12-20,22]. They decided on two methods, which were applied to a problem of credit scoring.

To support the first stage, the Kaufmann's Expertons Method is used [5,6]. Method applies interval pessimistic and optimistic estimates defined by the experts to reduce a possibly large number of projects considered for investment. The

knowledge is then condensed and compatibility levels on the set of possible risks for each project are built. For the further consideration only projects with the minor credit risk are selected.

At the second stage the chosen projects are compared and their ranking is made using the modified Possibilistic Discrimination Analysis Method. This method represents a possibilistic generalization of the known Fuzzy Discrimination Analysis [10] and is the modification of Possibilistic Discrimination Analysis Method previously proposed by the author's [20,22]. As a result, the possibility distribution is constructed on the set of all possible risks and on its basis, projects ranking is made.

The proposed technology includes the third stage, which is associated with an optimal distribution of investments among several projects.

In practice, often, investment is made into several projects, when each inquire different amounts. At the same time, the total investment amount is fixed. In such cases, it becomes necessary to decide which of the projects should receive investment. Taking into account the levels of possibility of projects crediting (obtained at the second stage) and also considering initial investment amount, the third stage applies bicriteria discrete optimization problem [3,4,16,21] for the most advantageous investment into several projects. This stage implies choosing the projects with the maximum possibility of crediting and the maximum profit. Description of this method see in Section II.

The research of the authors resulted in a new technology which was used in investment tender and supported the decision making. The article includes an example clearly illustrating the work of the proposed technology.

II. PROBLEM OF THE INVESTMENT'S OPTIMAL DISTRIBUTION

As the Expertons Method [5,6,8,20,22] is well-known and Possibilistic Discrimination Analysis Method was proposed by authors earlier [20,22], here we will describe only the method that supported the third stages of technology. Example given below, shows how Expertons Method and Possibilistic Discrimination Analysis Method are working.

Assume that after the completion of the second stage of projects' evaluation, there are n ranking projects with insignificant credit risks, and for each possible decision

(project) d_j the possibility level δ_j of its choice is calculated. We consider the issue of possible financing of the chosen projects in ℓ years.

Let's assume there are additional conditions for financing chosen projects. In particular, it is known that for financing of j -th project $j \in \{1, 2, \dots, n\}$ within i -th year $i \in \{1, 2, \dots, \ell\}$, a_{ij} monetary units are required, the profit received from implementation of j -th project constitutes c_j monetary units, and b_i monetary amount is allocated to finance chosen projects within i -th year.

In practice, the amount of funding, as a rule, is insufficient to satisfy all selected projects. Therefore, it is supposed that for at least one $i \in \{1, 2, \dots, \ell\}$ the inequality $\sum_{j=1}^n a_{ij} > b_i$ is true.

Considering the listed restrictions, the question arises as to which of the chosen projects should be financed to achieve maximum investment profits at the minimum risks. We offer the following solution of the problem.

If we introduce a Boolean variables $x_j, j \in \{1, 2, \dots, n\}$ by the following rule

$$x_j = \begin{cases} 1, & \text{if the } j\text{-th project is financed,} \\ 0, & \text{otherwise,} \end{cases} \quad (1)$$

we receive the following bicriteria Boolean linear programming problem:

$$\begin{cases} \max \sum_{j=1}^n \delta_j x_j, & (i) \\ \max \sum_{j=1}^n c_j x_j, & (ii) \\ \sum_{j=1}^n a_{ij} x_j \leq b_i, i = 1, \dots, \ell, & (iii) \\ x_j = 0 \vee 1, \end{cases} \quad (2)$$

where the criterion (i) represents the decision on selection of the group of projects with the maximum possibility, the criterion (ii) represents the decision on selection of the group of projects giving the maximum profit, while the conditions (iii) corresponds to the financial constraints.

The purpose of a problem (2): revealing and financing projects-group with the maximum possibility among those that give the maximum profit.

Thus, the objective functions will be:

$$1) f_1 = \max \sum_{j=1}^n \delta_j x_j - \text{choosing the projects-group with the maximum possibility;}$$

$$2) f_2 = \max \sum_{j=1}^n c_j x_j - \text{receiving maximum profit.}$$

To solve this problem we often apply the method developed by the authors for discrete possibilistic bicriteria problems [16,21].

Let X is the set of all Boolean vectors satisfying the conditions of the bicriteria optimization problem. Then by considering the scalar optimization problem

$$\lambda f_1 + (1-\lambda) f_2 \rightarrow \max, (x_1, x_2, \dots, x_n) \in X, \quad (3) \\ \lambda \in (0, 1),$$

with conditions (iii), where λ is a weighted parameter, we can find, in the general case, some Pareto optima [3,4,16,21].

Thus, the bicriteria discrete optimization problem can be solved by linear convolution of criteria.

III. APPLICATION OF MULTISTAGE DECISION MAKING FUZZY TECHNOLOGY

We developed the software package supporting decision making for optimal credit granting. The software was tested on concrete data. All the required information (results by expert commission consisting of 10 members) was provided by the "Bank of Georgia" and filtered according to our needs after the consultations with the managers of the Bank's crediting department.

A. Selection of Applicants with insignificant Risks based on Expertons Method

Processing the information with the Expertons Method allows for selecting only those applicants whose profile provides either insignificant credit risk.

Let's presume that the number of possible risk estimates for a given applicant, i.e. the number of possible decisions (crediting risks) equals to 4 $D = \{d_1, d_2, d_3, d_4\}$:

d_1 : crediting with an insignificant risk; d_2 : crediting with a low risk; d_3 : crediting with an average risk; d_4 : crediting with a high risk.

Assume that the members of the expert commission consider 4 possible decisions (i.e. the levels of credit granting risks for the concrete applicant) $d_j, j = 1, \dots, 4$. Instead of expressing their opinion by value $\alpha \in [0, 1]$, they provide confidence intervals which are included in the interval $[0, 1]$: $[a^*, a^*] \subset [0, 1]$, where a^* is the pessimistic level of given risk and a^* is the optimistic level of the risk. Such an approach gives an expert commission member a chance to be intellectually active and use his/her knowledge and experience when assigning the risk level.

Processing of the results of one of the applicants:

The aggregate table of experts' estimates has the following form:

TABLE 1. THE AGGREGATE TABLE OF EXPERTS' ESTIMATES

Experts <i>i</i>	Possible decisions d_j			
	d_1	d_2	d_3	d_4
1	[0.3,0.4]	[0.3,0.5]	[0.4,0.5]	[0.2,0.4]
2	[0.5,0.6]	[0.1,0.2]	[0.3,0.6]	0.5
3	[0.3,0.6]	[0.5,0.6]	0.5	[0.3,0.5]
4	0.9	[0.4,0.7]	[0.4,0.5]	[0.1,0.1]
5	[0.6,0.8]	[0.3, 4]	0.3	[0.5,0.7]
6	[0.2,0.7]	0.6	0.4	[0.2,0.3]
7	0.4	[0.8,1]	[0.6,0.8]	[0.1,0.2]
8	[0.4,0.6]	[0.4,0.6]	[0.3,0.7]	[0.4,0.5]
9	[0.5,0.7]	[0.2,0.4]	[0.3,0.4]	1
10	0.4	[0.7,0.8]	0.3	[0.5,0.6]

We consider statistics when to each possible decision $d \in D$ both the lower and the upper bounds of confidence intervals are assigned. The cumulative distribution law $F_*(\alpha, d)$ is then given by the expression $a_*^i(d)$, and, $F^*(\alpha, d)$ - by the expression $a_i^*(d)$, i is order of an expert. Thus, we receive

$$\forall d \in D, \forall \alpha \in [0,1]: \tilde{A}(d) = [F_*(\alpha, d), F^*(\alpha, d)],$$

where \tilde{A} denotes an interval experton.

Let us consider 11 α -cuts from 0 to 1, and for each of the possible decisions $d_j, j=1, \dots, 4$ calculate two statistics for each cut: one for the lower boundary of an interval and, the other, for the upper boundary. By extending these statistics to the set of levels $\{0, 0.1, 0.2, \dots, 0.9, 1\}$, we receive experton:

TABLE 2. EXPERTON

α -cut	Possible decisions			
	d_1	d_2	d_3	d_4
0	1	1	1	1
0.1	1	1	1	1
0.2	1	[0.9, 1]	1	[0.8, 0.9]
0.3	[0.9, 1]	[0.8, 0.9]	1	[0.6, 0.8]
0.4	[0.7, 1]	[0.6, 0.9]	[0.5, 0.8]	[0.5, 0.7]
0.5	[0.4, 0.7]	[0.4, 0.7]	[0.2, 0.6]	[0.4, 0.6]
0.6	[0.2, 0.7]	[0.3, 0.6]	[0.1, 0.3]	[0.1, 0.3]
0.7	[0.1, 0.4]	[0.2, 0.3]	[0, 0.2]	[0.1, 0.2]
0.8	[0.1, 0.2]	[0.1, 0.2]	[0, 0.1]	[0.1, 0.1]
0.9	[0.1, 0.1]	[0, 0.1]	0	[0.1, 0.1]
1	0	[0, 0.1]	0	[0.1, 0.1]

An experton \tilde{A} is then transformed as follows [6]:

- an averaged experton is calculated by taking a mean arithmetic value of each interval boundaries;
- the averaged experton is reduced to a possibility distribution $\pi_j = pos(\{d_j\})$, $j=1, \dots, 4$ on decisions set $D = \{d_1, d_2, d_3, d_4\}$ by taking mean values of all α -cuts;
- if necessary, a nonfuzzy set, the closest to the fuzzy one, is found.

After calculating the averaged experton and the mean values for each d_j on $D = \{d_1, d_2, d_3, d_4\}$, we receive the possibility distribution $\pi = \{\pi_1, \pi_2, \pi_3, \pi_4\}$ on D of identified risks of the considered applicant:

$$\{d_1/0.5727, d_2/0.55, d_3/0.4909, d_4/0.4818\}$$

To receive unique decision we apply the defuzzification for choosing the risk with principle of a maximum possibility: $\pi_{j_0} = \max_{j=1, \dots, 4} \pi(d_j)$.

This means that in accordance with the common opinion of the experts the experton gives preference to the decision (risk) d_1 , i.e. considered applicant has an insignificant crediting risk. Such a decision, of course, permits the applicant to participate at the second stage of the competition.

B. Ranking of the chosen Projects using the modified Possibilistic Discrimination Analysis Method

If at the first stage the Expertons Method based selection is carried out according to the classes of crediting risks, the next stage chooses from the number of the selected candidates by evaluating certain factors characteristic to these candidates. The possibilities of choice of candidates are evaluated by the Possibilistic Discrimination Analysis Method [20,22]. In our tender, after processing data by the Expertons method, for further consideration only six applicants remained from thirty-four possible candidates considered at the first stage.

Let us determine main $\omega_k, k=1,2, \dots, 9$ factors, by which all of the expert commission members will score the candidate juridical person seeking the credit.

The following factors influencing the decisions will be considered [12,20]:

ω_1 : business profitability; ω_2 : purpose of the credit; ω_3 : pledge guaranteeing repayment of the credit; ω_4 : credit amount (monetary value); ω_5 : payment of interest; ω_6 : credit granting date; ω_7 : credit repayment date; ω_8 : monthly payment of a portion of the principal and accrued interest (repayment scheme); ω_9 : percent ratio of the pledge to the credit monetary amount.

In our case, the value f_{ij} will designate the fraction of the experts who supported d_i decision when ω_j factor is present. (see [22], formula (1)). Thus, the expert commission consists of 10 members, the factors to be evaluated are $\omega_k, k=1,2, \dots, 9$, and, after the preliminary selection, the number of competitors equals to 6 ($d_j, j=1,2, \dots, 6$).

Suppose that the aggregate table f_{ij} looks like:

TABLE 3. THE AGGREGATE TABLE OF f_{ij} VALUES

Ω	D					
	d_1	d_2	d_3	d_4	d_5	d_6
ω_1	0.6	0.5	0.7	0.1	0.3	0.4
ω_2	0.4	0.8	0.1	0.5	0.2	0.3
ω_3	0.1	0.5	0.2	0.4	0.6	0.3
ω_4	0.4	0.3	0.6	0.1	0.2	0.7
ω_5	0.5	0.4	0.8	0.3	0.7	0.6
ω_6	0.6	0.1	0.3	0.4	0.5	0.2
ω_7	0.3	0.3	0.2	0.4	0.4	0.2
ω_8	0.5	0.2	0.3	0.6	0.4	0.3
ω_9	0.2	0.6	0.4	0.5	0.8	0.2

Firstly we calculate the table of π_j^i conditional possibilistic distribution (see [22], formula (2)):

TABLE 4. THE TABLE OF π_j^i CONDITIONAL POSSIBILISTIC DISTRIBUTION

Ω	D					
	d_1	d_2	d_3	d_4	d_5	d_6
ω_1	0.86	0.71	1.00	0.14	0.43	0.57
ω_2	0.50	1.00	0.13	0.63	0.25	0.38
ω_3	0.17	0.83	0.33	0.67	1.00	0.50
ω_4	0.57	0.43	0.86	0.14	0.29	1.00
ω_5	0.63	0.50	1.00	0.38	0.88	0.75
ω_6	1.00	0.17	0.50	0.67	0.83	0.33
ω_7	0.75	0.75	0.50	1.00	1.00	0.50
ω_8	0.83	0.33	0.50	1.00	0.67	0.50
ω_9	0.25	0.75	0.50	0.63	1.00	0.25

By converting it to the table of conditional probabilistic distribution p_j^i (see [22], formula (3)), we receive:

TABLE 5. THE TABLE OF CONDITIONAL PROBABILISTIC DISTRIBUTION p_j^i

Ω	D					
	d_1	d_2	d_3	d_4	d_5	d_6
ω_1	0.236	0.164	0.379	0.024	0.081	0.117
ω_2	0.119	0.556	0.021	0.181	0.046	0.077
ω_3	0.028	0.242	0.061	0.158	0.408	0.103
ω_4	0.136	0.088	0.279	0.024	0.052	0.421
ω_5	0.119	0.087	0.348	0.063	0.223	0.160
ω_6	0.408	0.028	0.103	0.158	0.242	0.061
ω_7	0.146	0.146	0.083	0.271	0.271	0.083
ω_8	0.228	0.056	0.089	0.394	0.144	0.089
ω_9	0.042	0.208	0.104	0.146	0.458	0.042

Further, to calculate the tables of positive and negative discriminations, we take the values $\alpha_1 = 0.25$, $\alpha_2 = 0.75$ (see [22], formulas (4)). As a result, we receive the table of positive and negative discriminations:

TABLE 7. THE TABLE OF POSITIVE AND NEGATIVE DISCRIMINATIONS

	D					
	d_1	d_2	d_3	d_4	d_5	d_6
p_{ij}	0.603	0.472	0.789	0.205	0.288	0.369
	0.448	0.876	0.207	0.569	0.270	0.350
	0.207	0.610	0.276	0.473	0.806	0.364
	0.457	0.355	0.641	0.206	0.272	0.811
	0.361	0.279	0.77	0.214	0.591	0.463
	0.806	0.207	0.364	0.473	0.610	0.276
	0.386	0.386	0.221	0.644	0.644	0.221
	0.599	0.213	0.285	0.799	0.454	0.285
	0.211	0.584	0.368	0.462	0.833	0.211
n_{ij}	0.218	0.262	0.175	0.469	0.366	0.311
	0.294	0.171	0.462	0.255	0.394	0.339
	0.463	0.221	0.385	0.269	0.174	0.323
	0.287	0.335	0.215	0.464	0.393	0.173
	0.304	0.360	0.175	0.428	0.216	0.258
	0.174	0.463	0.323	0.270	0.221	0.385
	0.274	0.274	0.403	0.184	0.184	0.403
	0.222	0.433	0.351	0.174	0.272	0.351
	0.441	0.236	0.319	0.276	0.173	0.441

We proceed with calculating π_j and ν_j representing the weighted average values of positive and negative discriminations for the j -th applicant (see [22], formula (5)).

Taking the coefficient value equal to $\beta = 0.95$ (see [22], formula (6)), we determine the possibility distribution on $D = \{d_1, d_2, \dots, d_8\}$:

TABLE 8. THE POSITIVE AND NEGATIVE DISCRIMINATION'S WEIGHTED AVERAGE VALUES AND THE POSSIBILITY DISTRIBUTION ON D

D	π_j	ν_j	δ_j
d_1	0.43735	0.30574	0.58143
d_2	0.46102	0.29780	0.59697
d_3	0.44837	0.30794	0.58582
d_4	0.43759	0.31749	0.57586
d_5	0.52554	0.26969	0.64230
d_6	0.37874	0.32766	0.54170

The Table 8 shows that ranking of considered projects according to possibility distribution δ_j is the following:

$$d_5 \succ d_2 \succ d_3 \succ d_1 \succ d_4 \succ d_6.$$

As can be seen from possibility distribution $\{\delta_j\}$, values of possibility for projects are sufficiently close. Therefore, if we need to invest into several projects, choosing from them will be difficult. In such cases, taking into account the amount of investment, proposed technology provides a third stage.

C. Optimal Investment into several Projects

Based on the possibility distribution $\{\delta_j\}_{j=1}^6$ of the chosen projects, obtained at the second stage of technology and on the given amount of investment, the third stage will deal with the bicriteria discrete optimization problem allowing for the most profitable investments into a number of projects.

For all six projects (the set of possible alternatives $D = \{d_1, d_2, \dots, d_6\}$), selected at the technology's second stage after processing them with Possibilistic Discrimination Analysis Method, we obtained possibility distribution:

$$\{d_1/0.58143, d_2/0.59697, d_3/0.58582, d_4/0.57586, d_5/0.64230, d_6/0.54170\}. \quad (4)$$

Bank considers an investment that totals to \$ 120 million over three years ($i = 1, 2, 3$), \$ 40 million a year ($b_i = 40$).

The values a_{ij} of investments, that are required for j -th project in i -th year, as well as the magnitudes c_j of profits from the realization of j -th project during three years are shown in the following table:

TABLE 9. THE VALUES OF a_{ij} AND c_j

Years	Projects						
	d_1	d_2	d_3	d_4	d_5	d_6	
a_{ij}	1	5	10	6	7	15	8
	2	2	10	6	8	15	7
	3	5	11	3	8	10	10
c_j	15	35	15	25	40	30	

Using formula (3) and the information given in (4) and Table 9, we solve the problem (2)-(3) taking for value $\lambda = 0.5$. As a result, we obtain the following set of Boolean variables $\{0, 1, 0, 1, 1, 1\}$.

This means that only four projects - d_2, d_4, d_5, d_6 - receive the credit.

At the same time, investment over the years amounted as \$ 40 million in the first and second years, \$ 39 million in the third year will bring the bank a total profit of \$ 130 million in three years.

IV. CONCLUSION

We developed the technology of experts information processing and synthesis. The technology is the combination of Kaufmann's expertons method, modified possibilistic discrimination analysis method and bicriteria discrete optimization problem. Based on this technology we have developed software package for decision making which is used to identify investment projects with minimum risks and optimal investment in several projects.

The application and testing of the software was carried out based on the data provided by the "Bank of Georgia". The recommendations of the financial managers and of the expert commission of the Bank were taken into account. The results are illustrated in the example.

The proposed technology, based on an analysis of the expert data is especially important for developing countries that lack sufficient reliable statistical (historical) data on investment projects.

The technology provides experts with the opportunity to manifest intellectual activity of a high level. Securing the freedom of experts' subjective evaluations, the technology, however, allows for developing experts' joint decision on granting credits. The latter distinguishes our technology from the others.

REFERENCES

- [1]. Bellman, R.,E., Zadeh, L.,A., "Decision-Making in a Fuzzy Environment, *Management Science*, " 17(4), 1970, B-141-B-164.
- [2]. Dubois, D., and Prade, H., *Possibility Theory: An Approach to Computerized Processing of Uncertainty*. New-York: Plenum Press, 1988.
- [3]. Ehrgott, M., *Multicriteria optimization*. 2nd ed. Berlin Heidelberg: Springer, 2005.
- [4]. Hamdy, A.Taha, *Operations Research: An Introduction, (9th Edition)*. Upper Saddle River, N.J.: Pearson/Prentice Hall, 2011.
- [5]. Kaufmann, A., "Theory of expertons and fuzzy logic, " *Fuzzy Sets and Systems*, 28(3), 1988, 295-304.
- [6]. Kaufmann A., *Expert Appraisements and Counter-Appraisements with Experton Processes, Analysis and Management of Uncertainty: Theory and Applications*, North Holland, Amsterdam, 1992.
- [7]. Klir, G., J., and Wierman, M., J., *Uncertainty-Based Information. Elements of Generalized Information Theory. Second edition. Studies in Fuzziness and Soft Computing*, Physica-Verlag, Heidelberg, 1999.
- [8]. Khutsishvili, I., "The Combined Decision Making Technology based on the Statistical and Fuzzy Analysis and its Application in Forecast's Modeling," *WSEAS Transactions on Systems*, 8(7), 2009, 891-901.
- [9]. Liu, B., "Toward fuzzy optimization without mathematical ambiguity," *Fuzzy Optimization and Decision Making*, 1(1), 2002, 43-63.

- [10]. Norris, D., Pilsworth, B.,W., and Baldwin, J.F., "Medical Diagnosis from patient records - A method using fuzzy discrimination and connectivity analysis," *Fuzzy Sets and Systems*, 23, 1987, 73-87.
- [11]. Ruan, D., Kacprzyk, J., and Fedrizzi, M.,. *Soft Computing for Risk Evaluation and Management: Applications in Technology, Environment and Finance. Studies in Fuzziness and Soft Computing*. Heidelberg, New York: Physica-Verlag, 2001.
- [12]. Sirbiladze, G., Sikharulidze, A., and Korakhashvili, G., "Decision-making Aiding Fuzzy Informational Systems in Investments. Part I - Discrimination Analysis in Investment Projects," *Proceedings of Iv. Javakhishvili Tbilisi State University. Applied Mathematics and Computer Sciences* , 353 (22-23), 2003, 77-94.
- [13]. Sirbiladze, G., and Gachechiladze, T., "Restored fuzzy measures in expert decision-making." *Information Sciences*, 169 (1/2), 2005, 71-95.
- [14]. Sirbiladze, G., and Sikharulidze, A., Weighted Fuzzy Averages in Fuzzy Environment, Parts I, II, *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 11(2), 2003, 139-157, 159-172.
- [15]. Sirbiladze, G. "Decision Precising Technologies in Decision Making Systems," Plenary Speech. In: *N. Mastorakis, M. Demiralp, I. Rudas, C. A. Bulucea, L. Rogozea (Eds.), Proc. of the Applied Computing Conference 2009, ACC '09, Vouliagmeni, Athens, Greece, 2009*, pp. 27-28.
- [16]. Sirbiladze, G., Ghvaberidze, B., Latsabidze, T., and Matsaberidze, B., "Using a minimal fuzzy covering in decision-making problems," *Information Sciences*, 179 (12), 2009, 2022-2027.
- [17]. Sirbiladze, G., Sikharulidze, A., and Sirbiladze, N., "Generalized Weighted Fuzzy Expected Values in Uncertainty Environment," In: *L. A. Zadeh, J. Kacprzyk, N. Mastorakis, A. Kuri-Morales, P. Borne, L. Kazovsky (Eds.), Proc. 9th WSEAS Int. Conf. on Artificial Intelligence, Knowledge Engineering and Data Bases, AIKED '10, University of Cambridge, UK, 2010*, pp. 59-65.
- [18]. Sirbiladze, G., Fuzzy dynamic programming problem for extremal fuzzy dynamic systems. In: *W. A. Lodwick and J. Kacprzyk (Eds.) Fuzzy Optimization: Recent Developments and Applications, Studies in Fuzziness and Soft Computing*, 254, Heidelberg: Physica-Verlag, 231-270, 2010.
- [19]. Sirbiladze, G., "Fuzzy Identification Problem for Continuous Extremal Fuzzy Dynamic System," *Fuzzy Optimization and Decision Making*, 9 (3), 2010, 233-274.
- [20]. Sirbiladze G. and Khutsishvili, I., "Combined Decision Precision Fuzzy Technologies for Credit Risk Evaluations of Bank Investments," *The Third International Conference "Problems of Cybernetics and Informatics", PCI, Baku, Az., 2010*, 193-197.
- [21]. Sirbiladze, G., Sikharulidze, A., Ghvaberidze, B., and Matsaberidze, B., "Fuzzy-probabilistic Aggregations in the Discrete Covering Problem," *International Journal of General Systems*, 40 (2), 2011, 169 -196.
- [22]. Sirbiladze, G., Khutsishvili, I., and Dvalishvili, P., "Decision Precising Fuzzy Technology to Evaluate the Credit Risks of Investment Projects," *IEEE 10-th International Conference on Intelligent Systems Design and Applications (ISDA 2010), Cairo, Egypt, 2010*, 103-108.
- [23]. Yager, R.,R., "On the Evaluation of Uncertain Courses of Action," *Fuzzy Optimization and Decision Making*, 1(1), 2002, 13-41.
- [24]. Yager, R.,R., "Aggregation of ordinal information," *Fuzzy Optimization and Decision Making*, 6(3), 2007, 199-219.
- [25]. Yu, L., Wang, S., and Lai, K. K., "An intelligent-agent-based fuzzy group decision making model for financial multicriteria decision support: The case of credit scoring," *European Journal of Operational Research*, 195(3), 2009, 942-959.

Maximization of Firm's profit Taking into Account Quality Criteria and Shelf Life of a Product

Shorena Okujava
Georgian Technical University
77 Kostava st.,
Tbilisi, Georgia
shorenaokujava@yahoo.com

Abstract— main problem when product is produced in lots is: what is an optimal lot size, which will give the maximum profit, with respect to production and storage costs and losses related with decrement in product's quality that occurs persistently in time. Consequently, to maximize firms' profit it is necessary to produce products in lots that minimize all of the three expenditures. In this Article represented are considered and economical-mathematical model and corresponding algorithms to solve a given task.

Keywords - optimization, production in lots, production costs, quality coefficient, profit, oriented graph.

I. INTRODUCTION

One of the most important events, which the whole world is monitoring with great interest and fear, is the global economical crisis, which has completely changed the existing economical situation all over the world as well as has influenced the revaluation of many economical rules and points of view. Intensive economic activity, with which the modern world was just characterized, is significantly decelerated. Big and small business organizations are doing their best to maintain the existing level of stability in order not to find themselves in the long list of already bankrupted companies. Under such critical conditions, firms started an extremely violent competition with each other. Business organizations hardly manage to find and gather consumers, let alone to assimilate new consumer markets. Today the consumers, who are also under the influence of the economical crisis, better respond to the lower prices of products, and the most effective way to lower product price is using the resources more effectively. That is why the optimal allocation of resources, which always was one of the most important issues, gains even greater importance today. If in the past the optimal use of production resources was the effective way to maximize the organization's profits, today this is the most important prerequisite for firm's stability and continual existence. The issue of making optimal decisions appears no less important. Time has always been an important factor for business organizations' efficient activity, but today the economical crisis has made its own corrections to this process as well. To assure the company efficiency and success, its management must make optimal decisions and do so fast.

II. PROBLEM FORMULATION

Because of the issues stated above, the decision of including information technologies in the decision making process becomes ever more attractive and even essential. Using economical and mathematical models, software, and model-based algorithms will maximally increase the effectiveness of managerial decisions and will reduce the alternative analysis time. Inspired by the multiform cases, which have arisen in practice according to the characters of produced goods, the production processes can be grouped in such a way that the models made upon common principles can be collaborated for large groups of production processes. Using these models will help analyze different variants and make decisions concerned with the future planning periods. In this article we discuss an economical mathematical model of optimal satisfaction of dynamic demand for products while maximizing producers' profit, taking into account the losses caused by diminishing quality of a product in time, production costs and storage costs. Using this developed model could be especially effective for small and mid-size businesses, which are producing products with limited shelf life, so that products' quality goes worth in time and it affects both the amount demanded and the price of a product. Discussed model is suitable for such a situation, where the production interval T is divided into numbers of subintervals or periods numbered $1, 2, \dots, n$. The demand for the product in each period is known, that is: $r_1, r_2, \dots, r_n, r_i \geq 0, i=1, 2, \dots, n$ and in i th period there can be produced the amount of product equal to $x, x_i \geq 0, i=1, 2, \dots, n$. If the supply of produced goods is greater than the demand for them, then for the future interval there appears the stock amounts of $y_i \geq 0, i=1, 2, \dots, n$. The size of the planning intervals and periods depends on the characteristics of concrete products and might be changed significantly. For example, for perishable goods the planning interval may be one day and periods may be hours. Similarly, the planning periods and intervals could be equal to, respectively, one week and one day, one month and one week, etc. As we already have mentioned we are discussing a case of maximizing producers' profit, which, according to economical theory, equals to: Total revenue mines total costs. In costs we are considering a sum of two types of costs: production and storage costs: $C(x)$ and $h(y)$. Fig. 1.a shows graphical relationship between production cost and a lot size, while Fig 1.b shows a relationship between storage cost and inventory size.

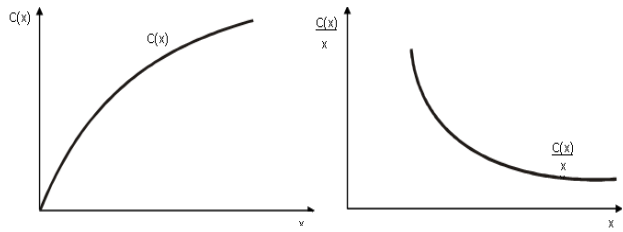


Fig. 1.a
 x – lot size
 $C(x)$ – cost of producing a lot of size x
 $C(x)/x$ - cost of production per unit of a product

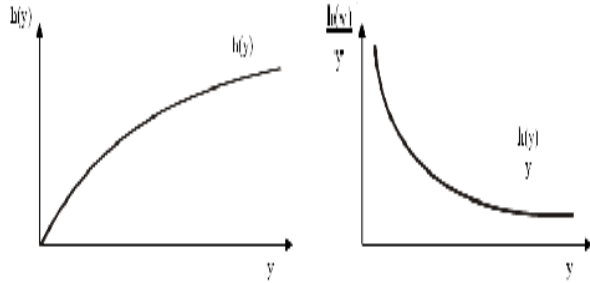


Fig. 1.b
 y – inventory size
 $h(y)$ – cost of storage of y
 $h(y)/y$ - cost of storage per unit of a product

We have discussed optimization cases by minimizing the production and/or storage costs. But as investigations have revealed, there is one more criteria – product quality – important enough to be included in the decision making process. Often decisions are made without foreseeing that quality criteria do not respond to modern needs of consumer market. So we would need to modify the case of optimization, taking into account the product quality criteria. The problem of optimization, including quality criteria can be formulated in different ways. We discuss the case when the company goal is to maximize its profit while minimizing the expected losses, which can be caused by a gradual deterioration of product quality. Note that this problem is solved with the condition that a producer is not limited in a capacity and can produce lots of any size. Such type of optimization is effective for the production situations, where organization is producing goods with expiration date, after which the product becomes useless for the consumer. Such products may be unsold even before the expiration date, because the consumer may choose similar yet fresher products. In general, product quality is reversely related to a shelf life of a product, and as time goes quality of a product decreases. According to how differently the quality index changes in time, products can be classified as follows:

1. Products with short shelf life;
2. Products with mid-length shelf life;

3. Products with long shelf life.

Products with short shelf life have short validation period and quality coefficient for such types of products diminishes very quickly. Fig.2.a shows the graphically how q quality coefficient decreases in time for these products.

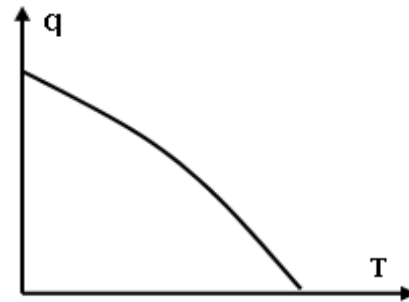


Fig. 2.a
 $T=1,2,\dots,n$. $q=[1,0]$

The validation period for mid-length shelf life products lasts from several months till years and quality coefficient diminishes stage by stage and reaches zero by the end of product's validation period. Fig 2.b shows such kind of relationship graphically:

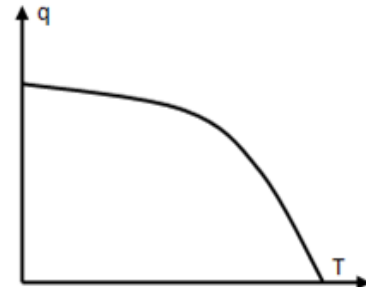


Fig. 2.b.
 $T=1,2,\dots,n$. $q=[1,0]$

Products with long shelf life do not have specified validation period, but their quality still diminishes slightly in time. Examples of such a product can be wooden materials, different kind of equipment and ect. Graphically quality coefficient for such kind of products changes as it is shown on Fig 2.c.

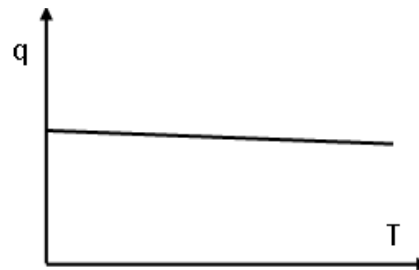


Fig. 2.c.
 $T=1,2,\dots,n. q=[1,0]$

In addition to above counted products, there are such products, for those quality index does not change during the whole validation period, but after expiration of this period, product immediately becomes useless. The relationship between quality coefficient and time period for these products is shown on Fig.2.d.

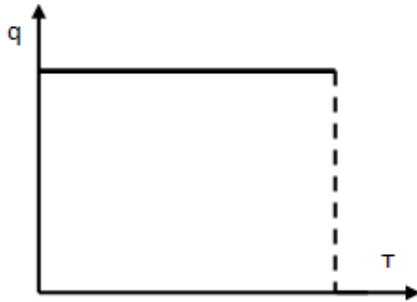


Fig. 2.d.
 $T=1,2,\dots,n. q=[1,0]$

III. PROBLEM SOLUTION

The problem of optimization, including quality criteria can be formulated in different ways. We discuss the case when the company goal is to maximize its profit while minimizing the expected losses, which can be caused by a gradual deterioration of product quality. Note that this problem is solved with the condition that a producer is not limited in a capacity and can produce lots of any size. Such type of optimization is effective for the production situations, where organization is producing goods with expiration date, after which the product becomes useless for the consumer.

Optimization with quality criteria must be solved using maximization process. So the problem of optimization can be formulated as follows:

Maximize:

$$\pi = \sum_{i=1}^n (r_i q_i p_i - (C_i(x_i) + h_i(y_i)))$$

where $y_i = \sum_{i=1}^n (x_i - r_i) \geq 0; \quad i=1,2,\dots,n$
 $r_i, x_i, y_i \geq 0, \quad x_i \cdot y_i = 0 \quad i = 1,2,\dots,n \quad q_i = [1,0]$

Where r_i is the demand for a product in planning interval, q_i is the coefficient of diminishing product quality during each interval, p is the market price of the product, $C_i(x)$ is the production cost of x_i lot, $h_i(x)$ is the storage cost of x_i lot, x_i is the lot size, and y_i is the amount of stored product during the planning interval. The goal of optimization is maximizing expected profit, taking into account production and storage costs as well as losses, concerned with the diminishing product quality during the time period. The algorithm of solving this problem is based on finding the largest path in oriental graph. We could use an algorithm of finding k number of the largest path in the graph, which gives several solutions, from which the decision maker must choose the optimal one, taking into

account the other criteria influencing the decision making process.

If the number of planning periods is equal to n , then the amount of nodes in the oriented graph will equal to $n+1$, where the last node shows the end of the planning period. The estimation of each arc in the graph is done with respect to the proper amount of profit. For example, the estimate of d_{13} arc (Fig.3) equals to the amount of profit, which will be received in the first period by producing the amount of product enough to satisfy the demand of the first two intervals.

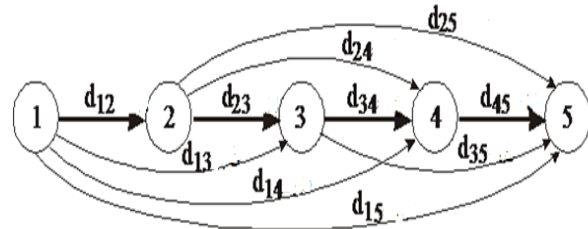


Fig.3
 Oriented graph

The software solution for this problem based on the algorithm discussed above is in the process of development for this moment. The results and conclusions will be discussed in further researches.

IV. CONCLUSION

As we have stressed out above, the problem of production process optimization with respect of different criteria is a very actual and important issue for business organizations. And using IT technologies and software solutions for these problems will help producers to optimize their production process, minimize costs and thus gain more profit.

REFERENCES

Lower. Optimal sequencing single machine subject to precedence constraints // Management science, v 19, N5, 1973
 [2] Nicholas C. Siropolis Small Business Management // Houghton mittlin company, Boston, 1986;
 [3] Baumol N. Economic Theory and Operations Analysis // 3d edition, Englewood Cliffs, N. J.:Prentice-Hall, 1972;
 [4] Kmenta J. Elements of Econometrics // 2d edition; New York; Macmillan; 1986;
 [5] Asatiani T, Lominadze N, The problem of minimization of the costs of production and holding inventory, Georgian Technical University, Scientific works, 1996;
 [6] Lominadze T, Gogichaishvili G, Problems of optimal sequencing of orders, Georgian Technical University, Scientific works, 1996;
 COMPUTING and COMPUTATIONAL INTELLIGENCE
 ISSN: 1790-5117 3

AN APPROACH TO SOLVING SOME MANAGEMENT PROBLEMS UNDER UNCERTAINTY

Teimuraz Tsabadze
dept. of computer sciences
Iv. Javakhishvili Tbilisi State University
Tbilisi, Georgia

Tengiz Tsamalashvili
dept. of social sciences and business
Tbilisi Teaching University "Gorgasali"
Tbilisi, Georgia

Abstract—this paper regards a one approach for solving some management problem by means of fuzzy sets theory. The control process is entirely based on the decision-making factor and is an outcome and content of the work of managers. Our objective is to solve most frequently occurring problems in situations which lack the previous experience and original information is weakly structured, fragmentary and/or incomplete. Here, as a rule, it is impossible to do without experts' evaluations which lead to the process of group decision-making. It is shown that set of experts' opinions presents a finite collection of fuzzy sets. A theoretical basis for aggregation of experts' fuzzy opinions established.

Keywords—management, Group decision-making, fuzzy set, coordination index, aggregation

I. INTRODUCTION

The management theory employs the scientific methods of synthesis and analysis to work out definite methods and recommendations for practical control, but it should be kept in mind that these methods and recommendations are not dogmas and they cannot be adopted as absolute rules. Therefore one of the important conditions of management effectiveness is that control methods be adequate to the external and internal medium in which a controlled object functions. Methods of various kinds of management may essentially differ from one another and the mechanical extension of a control method of one specific kind of management to another kind will not yield the desired results [1].

The control process is entirely based on the decision-making factor and is an outcome and content of the work of managers. Success or failure of controlled objects is a result of decisions made by managers. The components of the decision-making structure are purpose, techniques used to achieve results, evaluation criteria and rules of selection.

The main stages of the decision-making process are the receiving and preliminary processing of information and the decision-making procedure itself – this implies the formation of alternatives and their comparison, selection of a standard hypothesis or an action program, the construction and correction.

According to the character of problems, which arise in the course of activity of a controlled object, and methods used for decision-making, the following groups can be distinguished [2, 3]:

1. *Structured or, following the terminology of [4], programmed problems.* These are solutions which once made become the rules controlling all future actions. They are called structured (programmed) decisions and are part of daily activities of a controlled object, are permanently repeated and comply with the policy adopted by managers. They can be obtained by economical-mathematical methods and the algorithms, already developed, are available in most cases.

2. *Non-structured (non-programmed) problems.* A solution should be obtained in a non-standard situation with a lack of previous experience is characterized by incomplete knowledge about the main components and needs extraordinary approaches. In this area heuristic methods are of great use. Such solutions cause embarrassment in the manager because each time he has to seek for a procedure of their selection. This group includes in particular solutions for a concrete controlled object to get out of crisis, as well as solutions concerning a search for a strategy, determination of an amount of capital to be invested into new production, and so on.

3. *Scientifically substantiated standard problems solved by the statutory rules on solution selection.* Scientifically substantiated solutions are those which are adopted on the basis of scientific arguments. They are also called rational because they are supported by scientifically proven schemes incorporating various models, comparisons of different variants and so on. Obtaining such a solution needs much time, the process is frequently iterative, but enables the manager to find a solution.

4. *Intuitive problems and problems based on opinions.* As different from scientifically substantiated solutions, intuitive solutions are taken promptly, by feeling that one must act this way and not differently. Such solutions are based on personal experience.

It is obvious that managers encounter big difficulties when they are confronted with non-structured problems (item 2 in

the above list), which frequently have to be solved by heuristic methods.

Heuristic methods are meant to reveal, process and regularize rules, mechanisms and methods of anticipation, new buildings designing and purposeful actions on the basis of previous experience. Daily practice convincingly shows that as compared with the preceding century, processes occurring in the society in the 21st century are characterized by higher dynamism – this gives rise to a lot of new, hitherto unknown, problems. Since in the modern world the number of interconnected factors influencing many spheres of man’s activity steadily grows, the difficulty of making correct management decisions is obvious. No decision can be effective if from the very start it does not incorporate criteria of its realization [1].

In the present paper, we focus the attention on “the weakest link” – the non-structured process of making management decisions. Our objective is to solve most frequently occurring problems in situations which lack the previous experience and original information is incomplete. In our opinion confirmed by experience, after the detailed analysis of the management situation the next important step is to find input data to be used in making a final decision. Here, as a rule, we cannot do without experts’ evaluations which lead to the process of group decision-making. In this situation, the manager (managers) has (have) to solve the problem of alternatives aggregation.

We propose our concept of the solution of this problem.

II. ESSENTIAL NOTIONS AND THEORETICAL BACKGROUND

Managers (like all people) think in uncertain categories and are not guided only by the “yes” or “no” principle. For example, if the question is asked “Is this project profitable?”, classical mathematics answers either “yes” or “no”, while the relatively young fuzzy sets theory gives one a chance to model such categories as “unprofitable”, “not quite profitable”, “profitable”, “more profitable”, “very profitable” and so on.

Fuzzy sets theory made an invaluable contribution to the development of new information-control technologies. It is a powerful instrument of analysis of weakly structured, fragmentary, incomplete and fuzzy information. This instrument proved its high efficiency and potentiality when used in international and local organizations, corporations, various spheres of economy, business, sociology, politics and, generally, in new scientific and practical technologies.

The point is that in the absence of a general collection of data, i.e. of a sufficiently large initial database, even such well-tested instruments of modeling various situations as *probability theory* and *mathematical statistics* are not capable to fairly take into consideration incomplete and fuzzy information [5].

In 1965, American scientist L. Zadeh published the paper “Fuzzy Sets” [6], in which he drew the attention of the world scientific community to an absolutely novel direction in the development of mathematics and applied sciences. Instead of the classical membership function of some object in some set (0 or 1), Zadeh introduced the continuous membership area - [0; 1].

In the course of past decades, fuzzy sets theory made remarkable progress. Flexible and adequate methods were worked out for the solution of important applied problems with fuzzy initial inputs (data). The invaluable contribution of fuzzy sets theory to the development of decision-making methods is particularly noteworthy.

A fuzzy set A on a given (finite or arbitrary) universe X is defined as follows:

$$A = \{(x | \mu_A(x))\}, \quad x \in X, \mu_A(x) \in [0;1], \quad (1)$$

where $\mu_A(x)$ is the degree of membership of an element x to this fuzzy set. Hence it is clear that the transition from the membership to the non-membership of a universe element to some fuzzy set occurs not in a stepwise manner, but gradually running through the values from the interval [0; 1]. From (1) it follows that a fuzzy set is uniquely defined by its membership function.

$$\Psi(X) = \{\mu | \mu : X \rightarrow [0;1]\} \text{ - lattice of all fuzzy sets in } X$$

$$\emptyset \text{ - minimal element of } \Psi(X) : \mu_{\emptyset}(x) = 0 \quad \forall x \in X.$$

$$U \text{ - maximal element of } \Psi(X) : \mu_U(x) = b \quad \forall x \in X.$$

$$A = B \Leftrightarrow \mu_A(x) = \mu_B(x) \forall x \in X, \quad A, B \in \Psi(X).$$

$$A \subseteq B \Leftrightarrow \mu_A(x) \leq \mu_B(x) \forall x \in X, \quad A, B \in \Psi(X).$$

$$\text{Union of fuzzy sets } A \text{ and } B : \mu_{A \cup B}(x) = \max \{\mu_A(x), \mu_B(x)\} \forall x \in X.$$

$$\text{Intersection of fuzzy sets } A \text{ and } B : \mu_{A \cap B}(x) = \min \{\mu_A(x), \mu_B(x)\} \forall x \in X.$$

We say that the function $v: \Psi(X) \rightarrow \mathfrak{R}^+$ is *isotone estimation on* $\Psi(X)$ if

$$v(A \cup B) + v(A \cap B) = v(A) + v(B)$$

and

$$A \subseteq B \Rightarrow v(A) \leq v(B).$$

We say that the isotone estimation v is *continuous* if for each $a \in [v(\emptyset); v(U)]$ there exists $A \in \Psi(X)$ such that $v(A) = a$. The isotone estimation v determines the following *metric* on $\Psi(X)$:

$$\rho(A, B) = v(A \cup B) - v(A \cap B). \quad (2)$$

$\Psi(X)$ with isotone estimation v and metric (2) is called the *metric lattice* of fuzzy set.

In [7], based on the axiomatic approach, the notion of a coordination index of a finite collection of fuzzy sets defined

on the metric lattice of fuzzy sets is introduced. Necessary and sufficient conditions of the uniqueness of this coordination index are established. Putting aside the questions of completeness and independence of the introduced axioms, it is shown by constructing an appropriate example that this system of axioms is consistent.

Definition 1[7]. In the metric lattice the fuzzy set A^* is the representative of the finite collection of fuzzy sets $\{A_j\}$, $j = \overline{1, m}$, $m=2,3,\dots$, if

$$\sum_{j=1}^m \rho(A^*, A_j) \leq \sum_{j=1}^m \rho(B, A_j), \quad \forall B \in \Psi(X).$$

Definition 2[7]. The finite collection of fuzzy sets $\{A'_j\}$ is a regulation of the finite collection of fuzzy sets $\{A_j\}$ if for each $x \in X$ the finite sets $\{\mu_{A'_j}(x)\}$ and $\{\mu_{A_j}(x)\}$ are equal and $\mu_{A'_1}(x) \leq \mu_{A'_2}(x) \leq \dots \leq \mu_{A'_m}(x)$, $j = \overline{1, m}$, $m=2,3,\dots$

Theorem 1[7]. In the metric lattice of fuzzy sets the representative A^* of the finite collection of fuzzy sets $\{A_j\}$, $j = \overline{1, m}$, $m=2,3,\dots$ is determined in the following way:

$$A'_{[m/2]} \subseteq A^* \subseteq A'_{[(m+1)/2]+1},$$

Here and in the sequel, the square brackets denote an integer part of the number.

Theorem 2[7]. In the metric lattice of fuzzy sets the functional $S\{A_j\}$ is a coordination index of the finite collection of fuzzy sets $\{A_j\}$, $j = \overline{1, m}$, $m=2,3,\dots$ if

$$S\{A_j\} = q(\rho(\emptyset, U) - [(2m+1)/4]^{-1} \times \sum_{j=1}^m \rho(A^*, A_j)), \quad q > 0.$$

Moreover, if the isotone estimation v is continuous, this representation is unique.

For simplicity, let us give a particular discrete modification of the coordination index $S\{A_j\}$ of the finite collection of fuzzy sets $\{A_j\}$, $j = \overline{1, m}$, $m = 2, 3, \dots$, which is defined on the finite universe $X = \{x_1, x_2, \dots, x_N\}$, $N = 1, 2, \dots$:

$$S\{A_j\} = q(N - [(2m+1)/4]^{-1} \sum_{j=1}^m \sum_{i=1}^N |\mu_{A^*}(x_i) - \mu_{A_j}(x_i)|), \quad q > 0. \quad (3)$$

From (3.1.1) it obviously follows that

$$S_{\max}\{A_j\} = qN, \quad q \geq 0. \quad (4)$$

The method of fuzzy aggregation of expert evaluations is proposed in [8]. By the fundamental convention of the method, if experts have obtained the highest coordination value at some point of the universe, then potentially they may also obtain an analogous result at any other point of the universe. This "best" attempt of experts with a maximal coordination index is taken as the basis and all other attempts (evaluations at other points of the universe) are projected onto this basis. Next, a finite collection of one-element fuzzy sets is constructed at every point of the universe. Each of such collections is similar to the finite collection of one-element fuzzy sets at the point of maximal coordination and also the highest coordination index. Evaluations of each expert at all points of the universe are taken into consideration on equal terms.

In the end using the specific operator for fuzzy aggregation of expert evaluations and corresponding algorithm the result of group decision-making is obtained.

The specific operator for fuzzy aggregation of expert evaluations is as follows [8]:

$$\mu_{A^*} = \begin{cases} (\mu_{A'_{[m/2]}} + \mu_{A'_{[(m+3)/2]}}) / 2 & \text{if } \sum 1 = \sum 2 \\ \mu_{A'_{[m/2]}} + \frac{\sum 1}{\sum 1 + \sum 2} (\mu_{A'_{[(m+3)/2]}} - \mu_{A'_{[m/2]}}) & \text{otherwise} \end{cases}, \quad (5)$$

where

$\sum 1 = \sum_{j=1}^{[(m+1)/2]} \rho(A'_j, A'_{[m/2]})$, $\sum 2 = \sum_{j=[m/2]+1}^m \rho(A'_j, A'_{[(m+3)/2]})$ and $\{A'_j\}$ is the regulation of finite collection of fuzzy sets $\{A_j\}$, $j = \overline{1, m}$, $m = 2, 3, \dots$. The second specific operator is determined as follows [8]:

$$\mu_{A'_j} = \begin{cases} c + k \frac{v(B'_{[m/2]}) + v(B'_{[(m+3)/2]})}{2} & \text{if } \sum 3 = \sum 4 \\ c + k (v(B'_{[m/2]}) + \frac{\rho(B'_{[m/2]}, B'_{[(m+3)/2]}) \sum 3}{\sum 3 + \sum 4}) & \text{otherwise} \end{cases} \quad (6)$$

where

$$\sum 3 = \sum_{j=1}^{[(m+1)/2]} \rho(B'_j, B'_{[m/2]}), \quad \sum 4 = \sum_{j=[m/2]+1}^m \rho(B'_j, B'_{[(m+3)/2]})$$

$$\text{and } c = \frac{\sum_{l=1}^m (\mu_{B'_l}(x) - kv(B'_l))}{m}.$$

III. PROPOSED APPROACH

As mentioned in the preceding section, in uncertain situations it is advisable to make management decisions by using expert evaluations with the involvement of fuzzy sets

theory means. In this connection, there may arise very rational questions like “How will the manager interrogate experts?” and “How will the evaluations of experts who have no idea of fuzzy set theory be processed?”

Input data for the construction of a quantitative model of expert evaluations is the vector of targets of the considered project with a finite number of components. The vector is constructed by the manager (managers) after a careful study of the essence of the project. One of the most generalized examples of this vector is: {productivity, gain, costs}. It is obvious that each of these three hypothetical components consists of a finite set of parameters. The manager’s target is to obtain experts’ evaluations for each parameter and to make the final management decision.

The manager offers several experts to evaluate each component of this vector by any number from the interval [0; 1]. Suppose the vector consists of n components and there are m experts. Then a finite collection of fuzzy sets is formed, which consists of m sets having n elements in each $\{A_j\}$, $j = \overline{1, m}$, $m = 2, 3, \dots$, defined on the finite universe $X = \{x_1, x_2, \dots, x_N\}$, $N = 1, 2, \dots$.

Now the manager must construct by means of experts’ quantitative evaluations the resulting vector of the considered project. Let us describe how to do this.

It is clear that in each element of the universe we have a finite collection of one-element fuzzy sets, denote these collections as $\{B_j\}$. Now we introduce the procedure of fuzzy aggregation [8].

Algorithm

Step 0: Initialization: the finite collection of one-element fuzzy sets $\{B_j\}$, its regulation $\{B_j^i\}$, $j = \overline{1, m}$, $m = 2, 3, \dots$. Denote the result of the fuzzy aggregation in element x_i , $i = \overline{1, N}$ by $\mu(x_i)$.

Step 1: Compute the values of coordination indices of the finite collection of one-element fuzzy sets $\{B_j\}$ in each element x_i , $i = \overline{1, N}$ by (2.4). Denote these values by $S(x_1), S(x_2), \dots, S(x_N)$ respectively. Compute the value of S_{\max} by (2.5).

Step 2: Choose out of the set $\{S(x_i)\}$ such element S^* which is greater than or equal to any other elements except S_{\max} .

Step 3: Do Step 4 for $i = \overline{1, N}$.

Step 4: Compute $\Delta = S^* - S(x_i)$:

If $\Delta < 0$ then $\mu(x_i) = \mu_{B_j}(x_i)$;

If $\Delta = 0$ then compute the value of $\mu(x_i)$ by (5);

If $\Delta > 0$ then compute the value of k_i from the equation $S^* = k_i S(x_i) + (1 - k_i) S_{\max}$ and the value of $\mu(x_i)$ by (6).

Step 5: Representation is $\{\mu(x_1), \mu(x_2), \dots, \mu(x_N)\}$.

After obtaining the result of aggregation of fuzzy evaluations corresponding to the opinions of experts, the manager receives a good recommendation on making a final decision.

IV. CONCLUSION

In this paper we have stated the problem in general terms and, in conclusion, give a general example of how the presented material in this paper can be used by the manager. Suppose he considers some project represented by the vector of targets. It is necessary to select suitable candidates for a team which will realize the project. The task he faces is to determine coordination’s degree of evaluations of the components of the vector by pretenders to be project team members. The proposed quantitative model makes it possible to define a degree of maximal coordination among various pretenders and thus enables the manager to collect the most efficient team for the project realization.

REFERENCES

- [1] P. Drucker - The Effective Executive: The Definitive Guide to Getting the Right Things Done / Collins / 2006 / ISBN: 0060833459 / p. 208.
- [2] P. Schoderbek – Management Systems – 2nd ed – New York: Wiley, 1971, p. 124.
- [3] M. Meskon, M. Albert, F. Kheduri – Management – 3rd Edition: translation from English - M.: LLC “I. D. Williams”, 2007, p. 602.
- [4] H. Simon - The Sciences of the Artificial. The MIT Press. ISBN: 0-262-69191-4; 978-0-262-6 9191-8; 1996.
- [5] D. Dubois, H. Prade, Possibility theory, Plenum, New York, 1988.
- [6] L. A. Zadeh, Fuzzy sets, Inform. Control 8, No. 3 (1965), p.p.338–353.
- [7] T. Tsabadze, The coordination index of finite collection of fuzzy sets, Fuzzy Sets and Systems 107 (1999), p.p.177-185.
- [8] T. Tsabadze, A method for fuzzy aggregation based on grouped expert evaluations, Fuzzy Sets and Systems 157 (2006), p.p.1346-1361.

The model for predicting the competitiveness of science engineering products

Maslov A.V.

Yurga's Technological Institute (branch) National Research Tomsk Polytechnic University
Russia, 652050, t. Yurga, Kemerovo region, str. Leningradskaya, 26,
office tel.: 8 (384-51) 6-49-42, fax: 8 (384-51) 6-26-83
E-mail: mav00f@mail.ru

Abstract— A model for predicting the competitiveness of high-tech products according to changes in consumer preferences and improving the technical characteristics of products based on catastrophe theory (Abstract)

Keywords-component; forecasting of competitiveness; catastrophe theory; science engineering products

Leaders of innovative projects in their activities often have to deal with a situation of uncertainty or risk posed by the external environment. It may be, for example, receipt or loss of public procurement on competition, the presence or absence of material resource base, the sudden appearance of or withdrawal from market competition, changes in the requirements of potential investors or consumers, etc. Possible changes to be considered and modeled in a module in an open system compensate for negative factors up to the moment when the control subsystem is able to respond to them quickly, and efficiently generating tips. The dynamism of the environment leaves a progressive employer a minimum of time to analyze the situation and make a decision. If the progressive entrepreneur is not able to be active, learn from failures, to reduce losses to forecast levels of competitive alternatives and make them rational choice, it is inevitable for the time lag t_v in relation to the time of rational action. The objective of this module in an automated high-tech simulator determine the competitiveness of engineering products (NMP) - to predict the phase shift in the timing requirements of the consumer coupled with a change in the level of competitiveness of the NMP and revise the strategy of the enterprise (possibly complex) in accordance with the changing situation. In this direction, particularly useful are the elements of catastrophe theory [1,2,3], which allows us to develop models that take into account the important economic factors (profit, risk, sales volume, the level of costs, quality settings, the degree of novelty, etc.) that affect the competitiveness of products. Practice shows that the use of catastrophe theory, the possibility of such models may exceed the capabilities of scientific forecasting, such as technological.

Damages caused to the enterprise-producers during the slowing down of the reaction to changes in customer requirements to the parameters of NMP, depend on the ratio of the three time factors:

- t_d - the projection period; t_v - the time delay due to slow reaction to the threat; t_r - time to eliminate the threat and to improve the characteristics of products.

There are four possible situations that arise in the prediction:

1) the optimal situation where the time delay and time of removal of threats to coincide with the period of forecasting:

$$t_d = t_v + t_r;$$

2) If $t_v + t_r > t_d$, the situation corresponds to the fact that the reaction is delayed, and existing measures to stabilize not allow eliminate the risk before the start of "competitiveness crisis";

3) If $t_v > t_d$, extremely slow response and prognosis is associated with considerable difficulties;

4) If $t_r > t_d$, reaction time will come for the time horizon, and will not help reduce the time delay [3].

The original notation used in the model:

K_{nmp} - the competitiveness of high technology products and laundry facilities, score;

TEH_t - change TECH per unit time, conv. score;

PP - consumer of NMP preferences to TECH-estimates, conv. score;

R_n - the stiffness of communication (the degree of influence on the competitiveness TECH);

t - the interval of prediction.

Assumptions implicit in the model:

- Evaluation of competitiveness of production is directly dependent on the levels TECH provided by the manufacturer, and indirectly - from the PP formed a consumer at the time of the forecast.
- Evaluation TECH directly proportional to the PP.
- PP is a deviation of the actual levels of the TECH-optimal (rating levels of the hypothetical prototype NMP leader or equivalent).
- R_n ratio, taking into account the degree of influence on the competitiveness of TECH is equal to the value of the coefficient that takes into account the degree of influence PP on the levels of TECH. That is, in other words, how many levels of PP deviates from the base, the same amount should be changed corresponding to the level of TECH (optimality condition).

Competitiveness of the products on this TECH in a given time is determined by the formula

$$K_{nmp} = R_p \cdot TECH_t, \quad \alpha + \beta = \chi. \quad (1)$$

At time t due to the growth rate of technological progress is defined as a competitive

$$K_{nmp} = R_p \cdot TECH_t \cdot t \quad (1)$$

and the rate of change in competitiveness

$$dK_{nmp}/dt = R_p \cdot TECH. \quad (2)$$

Given the dependence of the PP TECH, according to which the greater levels of PP, the greater must be the value TECH products according to consumer preferences, and values of the coefficients equal to the degree of influence (optimality condition), we obtain

$$TECH = R_p \cdot PP. \quad (3)$$

Substituting (3) in (2), we have

$$dK_{nmp}/dt = R_p^2 \cdot PP. \quad (4)$$

Due to the increasing importance of PP, or the degree of influence on TECH R_p , and increased levels of PP, as the producer seeks primarily to satisfy the most important customer requirements. Conversely, increasing levels of PP, the manufacturer as the subject of dialogue, notes significant levels of PP higher estimates of R_p .

Based on the foregoing, we can write

$$PP = PP_0 \cdot (1 + \alpha \cdot R_p), \quad (5)$$

where α - degree of toughening PP [4].

Increased levels of PP, i.e., the risk of non-realization of products on the market, other things being equal, should be compensated by an increase in the level of competitiveness in these TECH (a necessary condition for holding positions in the market), i.e.

$$PP = PP_0 \cdot (1 + \alpha \cdot R_p - \beta \cdot R_p^2), \quad (6)$$

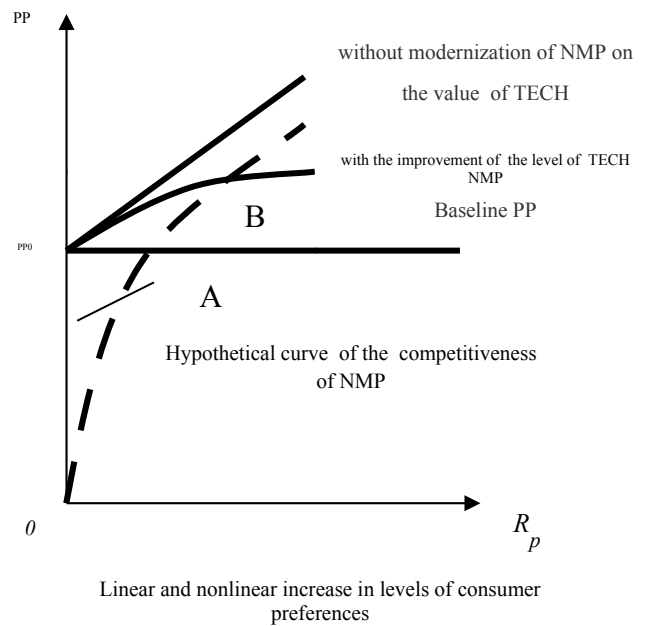
where β - the effectiveness of product-specific upgrade TECH (improving the value of production compared to the baseline).

In Fig. 1 shows the difference in the nature of growth in consumer preferences, depending on the measures taken to compensate for it with a significance level of R_p .

Points A and B, as shown in the figure characterize the situation where the needs are met in the NMP, respectively, at baseline and actual levels of NMP under the condition of constant (in the forecast period) to improve the characteristics of products.

Using the ratio of the simulation scenarios of the studied parameters, we can carry out the forecast rate of change in the competitiveness of the depending on the significance of the

coefficient of R_p , thereby prejudice, on what features will be provided with the largest margin of competitiveness.



Reducing the value of R_p reduces the level of competitiveness in the perception of a particular user (see Fig. 2).

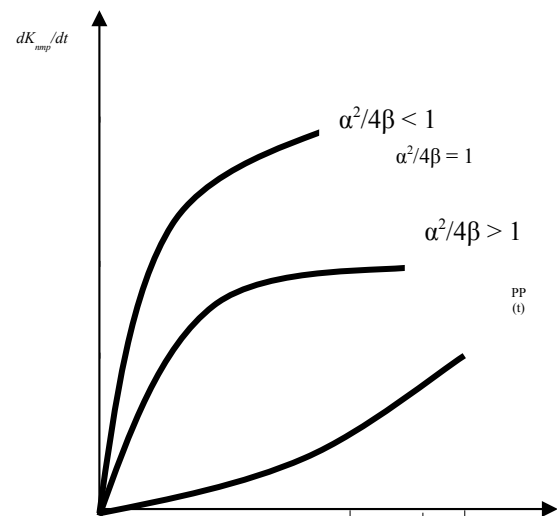


Figure 2. Hypothetical curves of the rate of change competitiveness of consumer preferences over time

If $\alpha^2/4\beta = 1$, when starting from a certain point, the rate of increase of competitiveness is reduced, if $\alpha^2/4\beta > 1$, the company accepts the risk of non-realization of products on the market due to inadequate response to the tougher requirements of consumers. To fully meet the demands of consumers and maintain market position, it is necessary to comply with a condition $\alpha^2/4\beta < 1$.

Due to slow the reaction products (the presence of time delays t_γ) and PP TECH products to the parameters do not correspond to the same time interval, or

$$\text{TECH} + t_\gamma \cdot d\text{TECH}/dt = \text{PP} R_p - \beta_0 \quad (7)$$

where β_0 - a necessary increase in the level of TECH with a possible increase of the PP and the degree of influence of R_p .

To construct the equation of the rate of change turn back the competitiveness of NMP (4), (6), (7). As a result, we obtain the following differential equation

$$dK_{\text{nmp}}/dt = \text{PP}_0 \cdot (R_p^2 + \alpha \cdot R_p^3 - \beta \cdot R_p^4) - (t_\gamma \cdot d\text{TECH}/dt + \beta_0) R_p \quad (8)$$

The right-hand side of (8) is a "competitive potential" products, including competitiveness, reducing the competitiveness of the resulting tightening of PP, the effectiveness of product-specific upgrade TECH, finally, necessary to compensate for the increase in TECH toughening of PP at a deceleration of the reaction products. The condition of maximum and minimum loss of competitiveness of products is the presence of a point at which the derivative of competitiveness will be zero. It follows

$$R_p^3 - (3\alpha/4\beta) \cdot R_p^2 - (1/2\beta) \cdot R_p + (t_\gamma \cdot d\text{TECH}/dt + \beta_0)/(4\beta \cdot \text{PP}_0) = 0, \quad (9)$$

For convenience and clarity, we present the resulting equation to canonical form, using the new control parameters a and b :

$$x^3 + ax + b = 0, \quad (10)$$

with

$$x = R_p - \alpha/4\beta;$$

$$a = -1/2\beta - 3\alpha^2/16\beta^2;$$

$$b = \alpha/8\beta^2 - \alpha^3/32\beta^3 + (t_\gamma \cdot d\text{TECH}/dt + \beta_0)/(4\beta \cdot \text{PP}_0).$$

The canonical equation of the form (10) in the practice of catastrophe theory describes the cusp catastrophe, which in [1,2,3] describes the set of potential functions that are displayed in an automated distributed systems [5].

Designed event model provides the following functions:

- Determination of the sensitivity of changes in TECH and PP are the basis for evaluating the attractiveness of products from a position of potential customers on the integral indicator of the competitiveness of K_{nmp} ;
- Operational processing and storage of statistical and expert information;
- Training of experts to make decisions quickly in crisis situations (situations of risk), and skills in the theory of catastrophes;
- Forecasting TECH and PP-based differential calculus;
- Determination of the optimal competitive products based on the minimum number of indicators;
- Generation of advice for decision-makers.

- [1] V. Arnold, Catastrophe Theory. Moscow: Moscow State University, 1990.
- [2] T. Poston, I. Stewart, Catastrophe Theory and Its Applications. Berlin: Springer-Verlag, 1980.
- [3] V. Tarasenko, Non-linear mathematical models and information systems in financial management. Tomsk: Univ. TPU, 1998.
- [4] O. Yakovets, Patterns of scientific and technological progress and its planned use. Moscow: Economics, 1989.
- [5] P. Ehlakov, E. Kornienko, V. Tarasenko, A. Shurygin, A. Kozintsev, N. Subbotin, «Distributed information technologies for collecting and analyzing information about the results of production, financial and institutional activities of the company», Journal of OWC. Moscow, vol.1, pp.100-102, January 1998.

Social CRM: a New Solution for Relationship with Bank's Customers

Mehrpooya Ahmadali Nejad
Faculty of Electrical, IT, and Computer Engineering
Qazvin Islamic Azad University
Qazvin, Iran
Mehrpooya@qiau.ac.ir

Seyyed Mohsen Hashemi
Assistant Professor, Computer Engineering Department
Science and Research Branch, IAU University
Tehran, Iran
Hashemi@srbiau.ac.ir

Abstract—Customer Relationship Management (CRM) has three main sections. First section is customers, other one is any service provider that provides services or products to customers that should be responsive for them and last one is relations between those two parts. Banking Customer Relationship Management follows this dominant too. Customers are in one side and the other side is the bank or financial institution. However in the current days due to the variety of interests, customer orientation and increasing the market competition, uncertainty and complexity, efficient and more innovative tools to meet that challenges are required. For this reasons and manifesting of new technologies in the virtual world such as Web 2.0 and social networks, we can improve quality and the satisfactory level of CRM.

Keywords—component; CRM; Social network ; Web 2.0; banking; Traditional CRM; Social CRM.

I. INTRODUCTION

CRM is a philosophy and a business strategy supported by a system and a technology designed to improve human interactions in a business environment. [1] This is a primitive definition of CRM. CRM in early days of usage was not implicated with Computer and information technology, but nowadays IT is a basis part of that. In these days and after implications of electronic CRM and customer orientation, a new perspective on customer relationship management has been created. Social CRM is coming from Social networks and Medias that can impact to CRM and make its definition changes.

In a Gartner report[2] 'estimated Though 2011, business-to-consumer (B2C) or business-to-business-to-consumer (B2B2C) enterprises will account for over 90% of spending on social CRM. During the next five years, community peer-to-peer support projects will replace Tier 1 contact center support in over 40% of the top 1,000 companies with a contact center. Though 2012, 90% of social CRM projects for sales organizations will focus on aiding prospecting and internal collaboration.

For a bank or financial institute, earn a vast source of data, relations, customers and other key elements with a great user friendly interface seems to be good opportunities.

In this paper we focus on a new concept of CRM in bank industry and we have discussed that social media helps the financial organizations to increase their Knowledge about customers and also help them to keeping customers.

The structure of the paper is as follows. In Sections 2 we have discussed about CRM and its evolution to social CRM. In section 3, is an overview on CRM systems in banking, its structure, Challenges and benefits. In Section 4, we discuss about designing Social CRM for banks and financial organizations, And in Section 5, we have concluded with future works.

II. CRM : FROM TRADITIONAL TO SOCIAL

A. CRM : in traditional way

CRM has developed as an approach based on maintaining Positive relationships with customers, increasing customer Loyalty, and expanding customer lifetime value. [3][4][5][6]

CRM's main functionalities are Marketing, Sales and Service. These three section can be important or not in different applications," Fig. 1". for example in bank industry service is the important part but in online stores each of parts are very important.

IT terms, CRM means an enterprise wide integration of technologies working together such as data warehouse, web site, intranet /extranet, phone support system, accounting, sales marketing and production. CRM has many similarities with Enterprise Resource Planning (ERP) where ERP can be considered back-office integration and CRM as front-office integration. A notable difference between ERP and CRM is that ERP can be implemented without CRM. However, CRM usually requires access to the back-office data that often happens through an office data that often happens through ERP-type integration. [7]

Customer Relationship management (CRM) originated from the contact management in 1980s, evolved into a development mode centered on customer through the customer care in 1990s, and nowadays becomes the focus of attention and development trend of global enterprise circles as a brand-new business philosophy.[8]

B. Social CRM : create a collaborative customer experience

Social CRM is a philosophy and a business strategy, supported by a technology platform, business rules, processes,

and social characteristics, designed to engage the customer in a collaborative conversation in order to provide mutually beneficial value in a trusted and transparent business environment. It's the company's response to the customer's ownership of the conversation.[1]

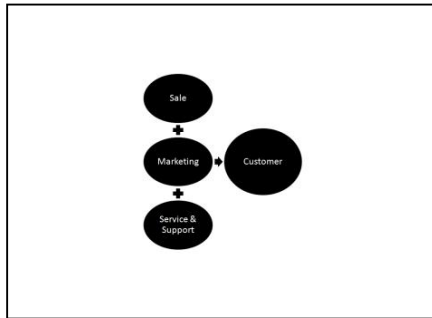


Figure 1. CRM's main functionalities

Before 2004 our exception from web and internet was a static environment that only for publication of news & reports and a one direction connecting. Web 2.0 technologies or what almost universally has been called social media is becoming accepted as an important digital tool used by firms partially with CRM.[9] A Social CRM system is designed for business to collaboratively manage business relationships and to create a collaborative customer experience initiatives to improve customer-organization relationships [10]. , "Fig. 2" illustrates customer collaborates.

In this situation Social customers are more knowledgeable, empowered and connected than ever before. prior to marketing any kind of purchasing decisions, customers now turn to peers and non-traditional industry influencers for answers through very public social networking platforms-more so than any other source for information gathering. Customer is a critical objective within any social business strategy. Social CRM strategies and technology offering should complement, but not replace traditional CRM software—the platform business have traditionally used to hold and analyze customer data, "Fig. 3". While many traditional CRM solutions are highly regarded and excellent at automating processes, managing the customer data and provide management reports to track sales, social CRM focuses on the relevant conversations taking place online and offline. [11]

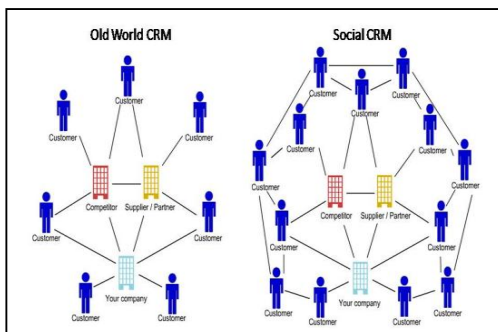


Figure 2. Evolution of CRM & Customer clbrates

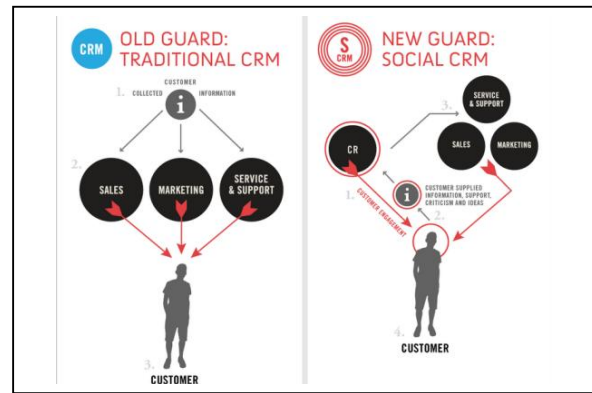


Figure 3. Comparing CRM Systems

Social CRM adds a new dimension and focus that works best on top of a solid foundation. This includes an easily accessible centralized customer database, keeping track of events and coordinating activities, and managing important sales and marketing processes. With the internal operational aspects being addressed by traditional CRM tools and strategies, the social layer aimed at engaging the wider Internet-based community can smoothly transition external conversations inward, continuing on the path towards a meaningful business relationship. [12] As mentioned, with the birth of the Web 2.0, social CRM was close to the time of its implementation. However wide variety of choices in Web 2.0 such as Social networks, wikis, media sharing, blogs was made a difficult situation to identify appropriate strategies to implement the Social CRM. But what does that make Social CRM more powerful, is the power of relationships between its users. Due to this fact, social networks have the greatest role in creating Social CRM, though other Web 2.0 technologies can also be involved to some extent.

C. Social CRM VS. Traditional CRM

As the "Table 1" illustrates, the focus of social CRM differs from traditional CRM in a few fundamental ways.

TABLE I. SOCIAL CRM VS. TRADITIONAL CRM

Traditional CRM	Social CRM
<i>Data driven</i>	<i>Content driven</i>
<i>Process centric</i>	<i>Conversation centric</i>
<i>Operationally focused</i>	<i>People/Community focused</i>
<i>PPT</i> <i>(People, Process, & Technology)</i>	<i>AAA</i> <i>(Automation, Analysis & Audacity)</i>

1) *Data driven vs. Content driven*: traditional CRM grew out of this need to store, track, and report on critical information about customers and prospects. Social CRM is growing out of a completely different need – the need to attract the attention of those using the Internet to find answers to business challenges they are trying to overcome.

2) *Process centric vs. Conversation centric*: Traditional customer relationship management is heavily focused on implementing and automating processes but in successful social CRM strategy, conversations are at the heart of it.

3) *Operationally focused vs. People/Community focused*: Whereas traditional CRM activity focused heavily on operational effectiveness and its impact – both internally and on the customer – social CRM is all about people and community.

4) *PPT (People, Process, & Technology) vs. AAA (Automation, Analysis, & Audacity)*: While any definition containing the importance of people, process, and technology captures the spirit of CRM, many were constructed before the Web became central to our lives. The philosophy behind social CRM is built upon a Web-powered foundation, and its impact on how we relate to each other. And with content being central to bringing people together, three other words become important to the equation – Automation, Analysis, and Audacity.[12]

D. Benefits of Social CRM

Social CRM applications as easy-to-use standalone applications that can be layered on the structured processes of existing CRM to help end users better leverage social networks, internal and external data and news feeds, and existing sales and marketing content.[13]

Social CRM applications share two key attributes. First, social CRM encourages organizations to share data by ceding control to a community through user-controlled mechanisms. The community is part of the process and decision making. Second, successful social CRM implementations provide clear benefit for an organization and its customers. Benefits to customers using social CRM applications are:

- Access to more trusted and independent information on products, services and organizations (including individuals inside an organization) through many-to-many participation.
- Personalization of interactions with an organization and products or services offer .
- Fulfilling emotional needs, such as self-esteem, respect, belonging and friendship. [2]

Social CRM system also can help organizations in following ways too:

1) *Single view of Online and Offline Constituent*: Social CRM allows organization to capture variety of online interactions such as e-mail activity and Web form donations automatically and Offline activities such as event participation or volunteer hours are also tracked.

2) *Provides Overall View of the System*: Social CRM serves not only its organization but its constituents also with a total view of their relationships with each other. This may include preferred methods of giving and visibility into how their contribution was used. This may also include tools that help organizations better manage and communicate with individuals in the same household, place of work, or alumni group.

3) *Comprehensive Knowledge and Oversight*: Social CRM provides the comprehensive knowledge and oversight of not just your constituents, but all of your fundraising efforts. Since most business processes are managed online, real-time reports such as integrated campaign activity statistics or customizable dashboards are always at fingertips.[13]

III. CRM IN TODAY'S COMMERCIAL BANKS

Since the early 1990s, computer, the Internet, and information technology have been merged to become a viable substitute for labor- and paper intensive financial processes between and across financial service providers. This has been seen in the widespread use of credit cards, debit cards, smart cards, and lending through CRM via the Internet. This computer-based transactional and informational exchange can be between each of financial service providers and also consumers, so it take place out of CRM and benefits of CRM to financial services providers and its customers.[14]

Evolution of CRM in banking industry can be categorized in two phases:

- Need of a place (system) where customer related information can be saved and extracted from here in raw format (as entered earlier) whenever needed.
- Need of analytics based on customer information available. This analytics is supposed to be used by marketing, sales and support team in such a manner that it should result in increased campaign responses, increased opportunity closure and reduced turnaround time for service requests.[15]

A. Systematic Structure of Banks' CRM

Banks' CRM refers to the promotion and application of CRM in the banking field of banks and falls into three levels: communication-level CRM, operation-level CRM and support-level CRM. "Fig4" shows these levels.

1) The communication-level CRM : the interface of CRM to interact, collect and output information with customers, including phone, fax, Internet, Email, wireless network, traditional counter, branch and others; customers can use various ways to get contact with customer service center, to retain the required information and service; thus, integration between advanced technology and all kinds of bank resources is effectively realized.

2) The operation-level CRM : made up of each sub module executing fundamental functions of CRM, including sales management module, marketing management module, customer service & support module, business intelligence management module, call center management module and E-commerce management module.

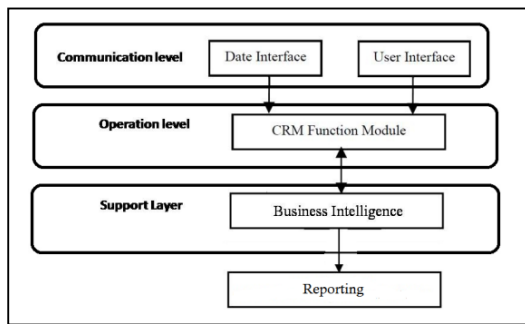


Figure 4. Systematic Structure of Bank's CRM

3) The support-level CRM The support-level CRM means data warehouse ,data mining technology that can consider this as a Business Intelligence system , operating system, network communication protocol and others, which are used in CRM, and is the foundation guaranteeing the normal operation of the whole CRM system. In the whole CRM structure, support-level CRM completes processing information accumulated from communication-level and operation-level, and then produces the analytical information based on the two. This information can be transferred to each functional module and front-consumer service system of the bank through systematic integration, and finally forms internal dynamic ,integrated consumer's analysis, management and service of the bank.[8]

B. Challenges with Traditional Banking CRM

Many analysts are of the view that CRM is currently seen as an administrative burden by the user. This is because of the fact that much of the CRM software is focused towards process automation and that does not necessarily provide incremental value back to the user. Sales people often see CRM as a reporting tool instead of sales generating tool and hence adoption wanes. There are two major challenges with traditional banking CRM system:

1) It was assumed that banking CRM system needs to own the data (of customers and prospects) to do analytics. Today, people may hesitate in providing their personal data to bank's call centre executive but they would like to enter the same information on social networks without even being asked to necessarily do so! In other words, customers prefer separate bank system, to interact, that space is not owned by banks.

2) Solution providers used to compete in banking sector based upon their product's features such as scalability and performance. Little emphasis was given to User Interface(UI).[15]

IV. LEVERAGING SOCIAL CRM FOR BANKS

Given what was said, Social CRM, a more complete relationship between organizations and customers to communicate. However the with Social CRM, Traditional CRM, should also continue to its functions, because the Social CRM, which entail information customers want to publish and Traditional CRM entail the information that banks need.

However, banks can use the additional information through Social channels to achieve a better strategy and decisions.

A. Responding to challenges of Traditional banking CRM

Web 2.0 websites allow users to do more than just retrieve information. It can be build on the interactive facilities of "Web 1.0" to provide "Network as platform" computing, allowing users to run software applications entirely through a browser[16]. Users can own the data on a Web 2.0 site and exercise control over that data [17]. These sites may have an "Architecture of participation" that encourages users to add value to the application as they use it [16]. Web 2.0 sites often feature a rich, user-friendly interface based on Ajax, Flex or similar rich media. The sites may also have social-networking aspects. [17]

for implementation of social CRM, we can Use two different approaches. First one is a private Social networking Sites –like bank mellat Customers Club[18]in Iran that however uses Web 2.0 opportunities but is not very user friendly for Social customers .

other Way that we focused in this paper is using a public Social networks like facebook, link din , tweeter and some places same these. In this way we have some problems too. Outside data may threat security issues but for earning more information from customers it can be worth to have some risks.

B. Cases of use & Benefits

Consider, the bank enters to a social media like social network. In social network about a bank's services, a debate arises among customers. After being exchanged comments from customers to reach results that enhance the quality of its services banks are very important. Here are two cases. The first is that the bank is passive in terms of passes and The social network is only used as a base for advertising. and another case that using a system of information exchanged and the desire to gain useful information from users. Even taken the information from the social network and existing user information in traditional systems can be integrated together.

For example, a bank intends to make awards based on the needs and tastes of its customers to donate. By the traditional system can be aware of the age, occupation, and its social conditions, but cannot be the best gift for him to consider, at this time that the Bank can best make use of customer information in social networks take. Perhaps at first glance it seems not very efficient, but should not forget that market competition is much more complex and more attention to any customer, whether or not he will result to customer loyalty and maintain.

Point to be cautious here is how the information obtained from different sources is used by the banking system. This information, if not used in correct manner may result in loss of trust in relationship between banks and their customers and hence the loss of business for banks. Conclusively, it makes sense to upgrade banking CRM with Web 2.0 capabilities and to assume that customer data may not be owned by CRM itself.[15]

C. Conceptual Systematic Structure of Social CRM for banks

The structure as a conceptual structure for Social CRM we have present, in the same Classification with traditional system structure and Falls into three levels: communication-level CRM, operation-level CRM and support-level CRM ,“Fig. 5”shows these levels..

1) The communication-level CRM : the social network as an interface of CRM to interact, collect and output information with customers. Customers have wide Choice to entering data

2) The operation-level CRM : The main operations that the system can be used to extract data from social networks. This extract can also be performed by human factors and also by data mining tools.

3) The support-level CRM : Here is also a business intelligence system that helps banks in decision making.

Our output in this way is some recommendations that can join with reports of traditional CRM .

V. CONCLUSION

Using Web 2.0 tools like social networks has created a situation that People with organizations and service providers have been closer. In this paper some of the situations mentioned in the banking and financial services. Despite the benefits of Social CRM, there are challenges such as inappropriate Rules in developing countries For Social networks and lack of trust and also technical complexity, in data mining and data Extraction are the most important of these. In our future works we Intend to eliminate this technical complexities.

Our goal to writing this paper is review and employing a comprehensive approach on Social CRM in the banks. We know that banks need their customers so each tools that earn more information about customers can improve their level of trust and services .

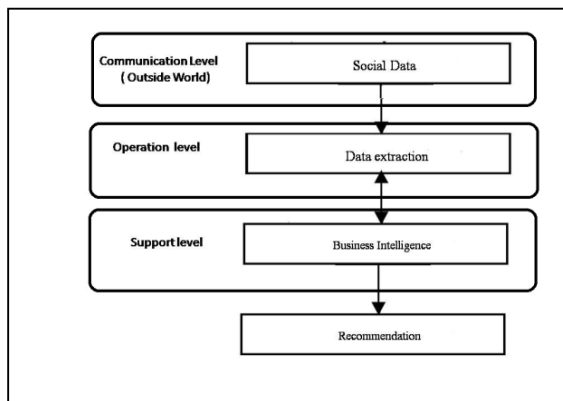


Figure 5. Systematic Structure of Banks' Social CRM

REFERENCES

- [1] P. Greenberg, " CRM at the Speed of Light, Fourth Edition: *Social CRM Strategies, Tools, and Techniques for Engaging Your Customers*," New York [u.a.]: McGraw-Hill, 2010, pp57-61 .
- [2] A.Sarner, E.Thompson, m.Dunne, J. Davies," Top Use Cases and Benefits for Successful Social CRM",Gartner, 2010
- [3] S.F. King, T.F. Burgess ,“Understanding success and failure in customer relationship management”, *Industrial Marketing Management* 37 ,2008 ,pp421–431.
- [4] Blattberg, R. C., & Deighton, J.. “Manage marketing by the customer equity test. *Harvard Business Review*”, 74(4), 1996, pp136-144.
- [5] Brassington, F., & Pettit, S. “Principles of Marketing (2nd edition)”, London: Prentice Hall, 2000.
- [6] Ahn, Y. J., Kim, K. S., & Han, S. K.. “On the design concepts for CRM systems”. *Industrial Management and Data Systems*, 103(5) ,2003, pp 324-331.
- [7] R. Bose, " Customer relationship management: key components for IT success" , *Industrial Management & Data Systems*, Vol. 102 Iss: 2 ,2002,pp. 89 – 97.
- [8] B.Fang, S.Ma ,” Data Mining Technology and its Application In CRM of Commercial Banks”, *First International Workshop on Database Technology and Applications*,2009.
- [9] S. S. Askool, K. Nakata,” Scoping Study to Identify Factors Influencing the Acceptance of Social CRM” , *IEEE ICMIT*,2010.
- [10] P. Greenberg, "CRM at the speed of light : essential customer strategies for the 21st century," New York [u.a.]: McGraw-Hill/Osborne, 2004.
- [11] Chess Media Group White paper;“Guide to Understanding Social CRM”, Chess Media Group, in Collaboration with Mitch Lieberman,June 2010.
- [12] B.Leary, “Social CRM: Customer Relationship Management in the Age of the Socially-Empowered Customer “,white paper of CRM Essentials, LLC,2008.
- [13] S.Mohan, E.Choi1, D.Min, “Conceptual Modeling of Enterprise Application System Using Social Networking and Web 2.0 “Social CRM System” ” , *International Conference on Convergence and Hybrid Information Technology* 2008.
- [14] M.Darajeh,M.Tahajod, “Benefits of e-CRM for Financial Services Providers”, *International Conference on Financial Theory and Engineering*, 2010.
- [15] Infosys White paper, “Social CRM- A Way to Innovate Banking CRM” Infosys, India , 2009.
- [16] T.O'Reilly (2005-09-30). What Is Web 2.0. O'Reilly Network.<http://www.oreillynet.com/pub/a/oreilly/tim/news/2005/09/30/what-is-web-20.html>
- [17] D.Hinchcliffe (2006-04-02). The State of Web 2.0. http://web2.socialcomputingmagazine.com/the_state_of_web_20.htm
- [18] <https://club.bankmellat.ir/>

Problems and Prospects of Electronic Shops Development in Georgia

David Zautashvili, Akaki Girgvliani
Akaki Tsereteli State University, Georgia

Abstract - There are considered in a given work the problems of Georgian electronic shops. There is presented the structure of electronic shops in Georgia by groups of goods. Also specific features of electronic business development in Georgia are emphasized. There are evaluated the prospects of Georgian electronic business associated with resolving the urgent problems.

I. INTRODUCTION

Under the conditions of new economy formation the role of Information Technology (IT) and Internet being the key driving force of innovations, economic growth and social transformations, had become increasingly important. Their development and ubiquitous use in the last decades of the XX century allowed changing fundamentally the centuries-long technologies of conducting the commercial operations by way of wide use of electronic data exchange when conducting these operations instead of traditional paper document circulation that laid the foundation to the rapid development of such phenomenon as electronic commerce.

Internet as a relatively new environment having extremely big opportunities in work with information had become an important element forming economic relations of modern business. IT opens to business in general and commerce in particular, new sources of income. Taking into account an influence of Internet on all aspects of economic life, it is possible to consider IT as a source of the creation of new electronic economy distinguished by extremely rapid growth, creation of new opportunities for industrial and business activity and an increase of number of workplaces.

It is obvious that ability of business and State structures to use the E-Business opportunities will become one of the key factors of competitive position of enterprises, companies and countries in the world economy of XXI century. The world indicators of the development of E-Business demonstrate this fact. There were over 2 billion internet users worldwide in 2011, and it is forecast this number will exceed 3 billion by 2015.

Online shoppers in the United States will spend \$327 billion in 2016, up 45% from \$226 billion this year and 62% from \$202 billion in 2011, according to a projection released by Forrester Research Inc. In 2016, e-retail will account for 9% of total retail sales, up from 7% in both 2012 and 2011, according to the report, "U.S. Online Retail

Forecast, 2011 to 2016," That represents a compound annual growth rate of 10.1% over the five-year forecast period. [1]

The EU online retail market generated close to \$163 billion 2010, reports MarketLine. Market growth is expected to slow to a yearly rate of just over 14% between 2010 and 2015 to exceed \$316 billion by the end of that five-year period. [2]

Considering the E-Business in Georgia, we must note that it counts relatively short history, only a few years. There are still too few electronic shops are functioning in Georgia at the moment, which face a lot of problems in their work. All these factors stipulate urgency of the studies of Georgian Internet-business development prospects and identification of the existing problems.

II. PRESENT SITUATION OF THE DEVELOPMENT OF E-BUSINESS AND E-SHOPS IN GEORGIA

Over the recent years, Internet in Georgia undergoes active evolution. According to Georgian National Communications Commission, the number of Internet users in Georgia in 2011 (including mobiles Internet) increased by 49% as opposed to indicators of preceding year.

The number of Internet users in 2011 has reached 1 637 905 people. Also, the used Internet traffic has increased: if in 2010 total Internet traffic amount was 180 terabyte, the same indicator in 2011 has doubled [3].

All that positively influences on electronic business development in Georgia.

Currently, there are existed in Georgia about 300 electronic shops. Unfortunately, very few Georgian Internet users are visiting almost one third of these electronic shops. [4], [5].

Structure of Georgian e-shops for each group of goods is given in figure 1. From the Table we see that the basic trends in online trade in Georgia are represented by sale of

electronic equipment (including computers, phones, and cameras), cars, motor vehicles, clothes, shoes and jewellery.

E-shops	Number of e-shops
Electronic equipment (incl. computer, phone, camera)	58
Clothes, shoes	24
Books	5
Jewellery	14
Electrical household appliance	9
Cosmetics	6
Children's products/toys	8
Product	6
Car, motor vehicle	65
Tools and do-it-yourself supplies	7
Gardening supplies	8
Other (Specify)	45

Fig. 1. Structure of Georgian e-shops for each group

There are some noticeable factors typical of electronic shops development at present stage in Georgia:

- all Georgian e-shops are narrowly-specialized, while over 10 categories of goods are selling on the leading shopping web-sites of USA;
- in most cases, speed in the operation of the web-sites as well as their operability leave a lot to be desired. This is conditioned by quality of equipment using by company and software package;
- poor design of the majority of web-sites impedes the potential client solicitation and retention. According to surveys, during the first eight seconds, a user has to catch sight of something useful for him on the site, otherwise he will simply move forward;

- the Internet companies almost do not advertising in traditional mass media. In USA, the advertisements of Internet companies are placed almost everywhere: in TV, newspapers and on the billboards;
- low service level, small number of professional delivery companies and logistics. Many Internet companies do not understand well enough that logistics governs in electronic commerce almost anything.

Analysis of electronic shops operating has revealed that as of today, the E-Business in Georgia faces many problems of various natures. Solution of these problems directly depends on electronic business development in Georgia.

Let's list the major problems of electronic business in Georgia:

- A small number of Internet users in Georgia. A number of Internet users in Georgia constitute only 30% of the entire population.
- Lack of professional specialists in E-Business, Internet-marketing and advertising.

Currently, all projects in the field of electronic business as a rule are implemented by IT-specialists, who do not possess required volume of economic knowledge. However, the electronic business development determines specific requirements for doing electronic business: they need in-depth knowledge and practical skills as in business field so in the field of new technologies. Unfortunately, the institutes of higher education in Georgia in practice do not prepare specialists in E-business field. There are no E-business subjects in syllabuses for preparing specialists in Information Technology and Economics (the exceptions are only Georgian Technical University and Akaki Tsereteli State University).

- Absence of normative and legal bases in electronic business

It is necessary to adopt in Georgia appropriate laws related to electronic business. On March 14, 2008 the Georgian legislative body adopts the law on Electronic Signatures and Electronic Documents. However, there are no adopted still other significant legislative documents such as: Law on Electronic Trade; Law on Electronic Deal, which themselves create that legal environment, wherein the electronic trade will be freely developed

- In practice there is no statistics on electronic business

Raise of the statistics on electronic business to high professional level will enable legislative power to rely on statistical materials and develop effective normative-legal framework, which would regulate legal relationships in this field. That will be problematic without organizing the adequate statistical monitoring. There no data in Georgia about visitors of Internet shops and studies of interests of

their users that is a matter of interest for specialists in electronic commerce.

- Low income level of population

It is necessary to strengthen the economic sector, ensure growth of income level and formation of middle class. Georgia with its indicator of GDP per capita, which was estimated at \$5,450, was ranked at 111th among 181 countries. This indicator is considerably lower than average world level, which equals \$10,700 at the moment.

- Security problem of financial transactions

Weak protectability of clients and storefronts from fraud and low distribution degree of credit cards among population are observed in Georgia at the moment.

Among subjective factors preventing electronic business development in Georgia we can emphasize the following ones: low confidence of potential buyers and psychological unwillingness of buyers to purchase goods via electronic trade.

III. CONCLUSION

There are much more problems and questions in Georgian sector of electronic business, than answers and ready schemes and solutions. Electronic Business in Georgia is appeared and described as a sphere of unsolved problems, while in USA and Europe, the description goes in the movement of experience and achievements.

We believe that very promising for small and medium Internet shops in Georgia should be the filling own niche on the basis of exclusivity and unique of goods.

We think that the “mobile commerce” (M-commerce) development should be a very promising way for Georgia. According to Georgian National Communications

Commission, the most growing interest of users in mobile Internet-services was observed in 2011. While in the last quarter of 2010 there were about 800 thousand users of Mobile Internet, in the last quarter of 2011 this indicator exceeded 1,2 million mark (50% higher).

In order to develop electronic business in Georgia it is necessary to ensure promotion of electronic business, change strategy of already existed e-shops and gain and increase confidence of users.

Only after taking these measures it is possible to expect that the e-shops with strong management system, structured framework, business-plan, correct legal system, entry strategy into market and marketing policy will appear in Georgia.

REFERENCES

- [1] “US Online Retail Forecast, 2011 To 2016” by Sucharita Milpuru. 27 Feb 2012
<http://www.forrester.com/US+Online+Retail+Forecast+2011+To+2016/fulltext/-/E-RES60672>
- [2] “e-Commerce Industry: Market Research Reports, Statistics and Analysis” April 2012
<http://www.reportlinker.com/ci02106/e-Commerce.html>
- [3] The Georgian National Communications Commission
http://www.gncc.ge/index.php?lang_id=ENG&sec_id=50623
- [4] Web site analyser
<http://www.wsa.ge>
- [5] Rating of populary Georgian sites
<http://www.top.ge>

Information Access in the Globalised World

Mariam Paposhvili
Iv.Javakishvili Tbilisi State University
Field of study: International Relations, 3rd year
Tbilisi, Georgia
mariampaposhvili@hotmail.com
Aleksandra Suladze

Iv.Javakishvili Tbilisi State University
Field of study: International Relations, 3rd year
Tbilisi, Georgia
sandrasuladze@yahoo.com
Tbilisi, 2012

Abstract - In this paper, focus is given to recent advances in technology and the process of globalization. It assesses the important aspects related to public access to information in the new world. A much wider significance is given to mass communication as it plays important role in the process of globalization. The major features of this process are introduced, discussing its multilateral form (economic, military, environmental globalization and etc). In addition positive as well as negative effects of globalization on mass communication are illustrated. The paper highlights the role of the media and states that in different cases the mass media can be used as a tool for achieving own purposes by government or different interest groups. The article examines different aspects of development of mass communication and pays attention to difficulties related to huge amount of information.

Introduction

The world has dramatically changed. In recent years we have witnessed great advances in technologies and their immense role in changing patterns of life. The world has become more interconnected by the process of globalization as states, societies, economics and cultures in different regions of the world are interdependent and increasingly integrated. Development of transport, new technologies give opportunity to understand more about the whole world. From the internet or TV we can get new information about other countries, people and their culture. We see how the life is going on in different societies. Globalization is a process of destroying borders between countries. Social relationship isn't disturbed by time or distance, if we needed two or more weeks for our letter to be delivered to someone, now by using of e-mail or Skype we may do it in a few seconds.

In the process of massive changes social media seems to be connected link among multiple actors and groups, on its own account media enhances the public access to information. How can we assess the consequences? The current processes clarify that information does not flow in a vacuum but in political space which is already occupied.¹ The

¹ Robert O. Keohane, Joseph S. Nye *Power and Interdependence in the information age*, Foreign Affairs vol. 77 No. 5 (Sep-Oct, 1998), p.84

media is often self-serving and used for propaganda purposes as well as for mass public control. Despite its negative reveal social media as communicative tool represents public interests and in certain circumstances its gave relevant power to people.

Aspects of globalization

The globalization is revealed in different dimensions therefore it lets us discuss its various aspects. Thus equally important forms of globalization stated by Robert O. Keohane and Joseph S. Nye are:

Economic globalization refers to increasing economic interdependence of national economies across the world through a rapid increase in cross-border movement of goods, service, technology and capital. It also involves the organization of the processes that are linked to these flows, such as the organization of low-wage production in Asia for the U.S and European markets.

Military globalization refers to long-distance networks of interdependence in which force, and the threat or promise of force, are employed. A good example of military globalization is the "balance of terror". In global sense new challenges are concerned with proliferation of global arms, nuclear weapons.

Environmental globalization refers to the long-distance transport of materials in the atmosphere or oceans, or of biological substances, that affect human health and well-being. Globalization transformed local environmental problems into international issue, it is hard to find a place which is being left untouched by major environmental problems such as the depletion of the ozone layer; the spread of the AIDS virus from west equatorial Africa around the world since the end of the 1970s.

Social and cultural globalization involves the movement of ideas, information, images, and people (who, of course, carry ideas and information with them). Examples include movement of religions or the diffusion of scientific knowledge. At its most profound level, social globalism affects the consciousness of individuals and their attitudes toward culture, politics, and personal identity. The global proliferation of communication carries ideas and currents across continents, sensitizing remote people to similar agendas and promoting mutual

intentions.²

The globalization makes sense of being part of “global village” and concerns all aspects of human existence. All abovementioned aspects are tightly linked to each other encouraged by technological progress and in this process role of media can be considered as a dominant. Overlooking the current process in economics, politics, and culture, and we see the media as essential in every phase. Indeed the evolution of global trade would be impossible without a flow of information on markets, prices, commodities, as well as the development of music, poetry, film, fiction, cuisine, and fashion would be hard to imagine. Globalization and media have proceeded together through time. Its dominance is caused by technological innovations and progress that is a driving factor in the process of globalization. Development in the early 1990s in computer technologies and telecommunications have caused extensive improvements in access to information and economic potential. Creating efficient and effective channels to exchange information, new technologies has become the catalyst for global integration. It’s an obvious fact that information now flows more freely and consequently media represents one of the most significant player in the process of globalization.

Media globalization is not a recent phenomenon.³ We can say that its development started in 1850 with invention of telegraph system which enabled people to become part of worldwide network. The opening cave paintings to papyrus to printing presses to television to Facebook, media have made globalization possible.⁴

In fact, history suggests a remarkable series of stages of globalization, each made possible by quite different technological advances. Specifically, building on Thomas Friedman’s (2005) formulation in *The World is Flat: A Brief History of the Twenty-First Century*, we put forward three stages:

Globalization 1.0 (1400 A.D. to WWI) with changes in transportation technology allowing the great explorers like Vasco de Gama (1460-1524) and Christopher Columbus (1451-1506) and culminated in steamships and airplanes

Globalization 2.0 (WWI to 2000) with changes in communication key, giving us telephones, fax, radio and TV, and e-mail.

Globalization 3.0 (2000 to now) with computing power key, allowing PCs (personal computers) to be linked by

fiber optics and the initiation of GRID computing (an emerging global, distributed parallel processing infrastructure)

Each stage has had more impact on the way people can work with information. Globalization 1.0 allowed mail packets—now dubbed “snail mail” packets—to be sent around the world in months and then days. Globalization 2.0 made communication electronic and cut the global circuit to seconds. Globalization 3.0 not only makes communications links quicker, but also makes them more complex, increasing the density of the web of connections.⁵

Positive effects of globalization on mass communication

UNESCO published in 1978 "The Declaration of Fundamental Principles Concerning the Contribution of Mass Media"; article V refers to the mass media effectively contributing to the strengthening of peace and international understanding, to the promotion of human rights, and to the establishment of a more just and equitable international economic order.⁶

Media is direct source of social sensibility. For example American civilians sitting in front of the TVs and watching what happens in another part of the world, that millions of people starve and suffer from poverty, increase the pressure on government to take appropriate measures to cope with.

The media in its positive sense is the process of building a virtual reality, its role encompasses increasing human solidarity , it demand government to be apparent and answerable to the people, to pay the government attention to the important public problems, regular updating of the memories of society, change acceptance, awareness in current

affairs and etc. In postindustrial society, information is a power factor - knowledge is power.

Critics on effects of globalization on mass communication

People have an access to too much source of information, which causes ambiguity and uncertainty as it makes different information difficult to filtrate. A plenitude of information leads to a poverty of attention. Attention becomes the scarce resource, and those who can distinguish valuable signals from white noise gain power.⁷

The researchers against globalization consider news as poisoning, polluting our brains, manipulating in order to inoculate us, as media consumers, subconscious ideas that are not our own. For this reason, the same researchers

² Robert O. Keohane, Joseph S. Nye, *What’s new? What’s Not? (And So What ?)* Foreign Policy, No. 118 , Spring, 2000 , pp.106-107

³ Nicoleta Munteanu, *Effects of Globalization on Mass*

Communication, Read Periodical, October 1, 2011
<http://www.readperiodicals.com/201110/2561867691.html>

⁴ Jack Lulu, *Globalization & Media, Global Village of Babel*, Rowman&Littlefield Publishers, USA,2012, p.5

⁵ George T. Duncan, Stephen F. Roehrig, *Reconciling Information Privacy and Information Access in a Globalized Technology Society*, Chapter 5.14, Carnegie Mellon University, USA, 2009, p.1824

⁶ Nicoleta Munteanu, ...

⁷ Robert O. Keohane, Joseph S. Nye, *Power and Interdependence in the Information Age*, Foreign Affairs, September/October 1998, p. 91

consider that is absolutely necessary to establish ecology of news, to sort real news from lies, to decontaminate the news we receive. Just as we can buy organic food less contaminated, we need biological news. The same authors insist that the news consumers should demand global owner media groups to show respect for the truth, because the news is legitimated only when are engaged in a search for truth.

Role of media- different aspects

The role of media cannot be unilaterally defined as it varies according to circumstances. In some cases the media can be used as a tool for government to maintain power and the control on the people, on the other hand it can be seen as an effective mean of political campaign for elections, however the media communications can play an important role in realization of public interests. To make our discussion credible we can examine some cases which highlight different characteristics of the role of the media.

Media as political campaigning tool

Enabled by pervasively accessible and immense communication techniques, social media subsequently change the

way of communication between government and public. Social media play important role in campaigning process, it is easy and profitable way to manage and make good and successful election campaign.

Three periods of campaigning are identified: **a newspaper stage**, (premodern campaign), **a television stage** (modern campaigning) and **a digital stage** (postmodern campaign).

⁸Hence, media technology enjoys a outstanding position as a force of change. Consequently, campaign process was implemented through the party controlled newspapers, and agitation, and aimed to mobilize voters. Campaign strategies were oriented on a single, coordinated national message that intended to attract voters across different social categories.

Effectiveness of social media as a campaigning tool seems for Barack Obama's election campaign. Barack Obama is the first "Social Media President" in history.⁹ After the 4th November 2008, when Barack Obama has finally entered the White House, there was no doubt worldwide, that he and his campaign team have changed the traditional way of political campaigns. Just as John F. Kennedy initially established the television to political campaign, Obama's presidential campaign changed the way political campaigns utilized the internet, specifically social media, for political purposes. Although, the effect of the internet has already been increasing during the previous

⁸ Rune Karlsen *Does new media technology drove election campaign change?* Information Polity 15 (2010) 215–225, p.216

⁹ Li Evans, *Barack Obama The First Social Media President?* Search Marketing Gurus, November 5, 2008
<http://searchmarketinggurus.com/2008/11/barack-obama-the-first-social-media-president.html>

presidential elections and they showed that Web sites are an important tool for direct communication between the candidates and their electorate and to mobilize fund-raising. In particular, the social media was used and nowadays, Facebook, Twitter and blogging seem to have important political campaigning meaning.

Another fact emphasis the role of social media and its impact on public during the elections regarding the Obama's presidential campaign is Oprah Winfrey's endorsement of Barack Obama.¹⁰ In total it is estimated that the endorsement was responsible for 1 015 559 votes.¹¹ Her influence through the media was noted with the conference "Global Oprah: Celebrity as Transnational Icon" by the Yale University.

Media and Arab Spring

Arab Spring is an example how social media can be used to set goals, show solidarity and organize mass demonstrations. Mohamed Bouazizi, the young merchant, set himself on fire in protest of the government in front of a municipal building in Tunisia, this became main discussing point on Facebook, Twitter, and YouTube in ways that inspired people to organize protests, criticize their governments. Social media encouraged democratic fervor spread across North Africa and the Middle East. Social media promoted the Arab Spring and take a power to put a human face on political oppression.

Governments themselves also recognized the power of opposition movements equipped with social media. In Tunisia, officials attempted to block Facebook and other social media sites and arrested bloggers and others who used social media to spread critical news about the government.¹²

Media under government control - case of North Korea

North Korea is one of the most strictly controlled states in the world. Press freedom doesn't exist in any sense as deviation from the official government is not tolerated. Use of personal computer in the country has increased significantly. Although North Koreans cannot access the global internet. It is accessible to only limited minority, who get information by illegal ways. The average North Korean today is far more aware of the outside world. Foreign media broadcasts are banned.

¹⁰ Oprah Winfrey is an American media proprietor, talk show host, actress, producer

¹¹ *How Bad Is It for Obama That Oprah Won't Campaign for Him?* The Atlantic April 2, 2012

<http://www.theatlantic.com/politics/archive/2012/04/how-bad-is-it-for-obama-that-oprah-wont-campaign-for-him/255357/>

¹² Philip N Howard, Aiden Duffy, Deen Freelon, Muzammil Hussain, Will Mari, and Marwa Mazaid 2011 *Opening Closed Regimes: What Was the Role of Social Media During the Arab Spring?* Project on Information Technology & Political Islam 2011, p.3

Although most urban households have radios and some have television sets, neither radios nor televisions can be tuned to anything other than official programming. Only some 10 percent of the radios and 30 percent of the televisions are in private households. Government control extends to artistic and academic circles, and visitors report that the primary function of plays, movies, books, and the performing arts is to contribute to the cult of personality surrounding Kim Il Sung.¹³ The government controls all cultural and media activities, forbids the public meetings, founding independent associations and organizations. As government maintain total power on the media uses it for propaganda and spread the misleading information according to own interest.

Conclusion

“However, no matter how you feel about the world, globalization has shaped your world. And globalization will shape your world. From the food you eat, to the cell phone you use, to the schools you attend, to the air you breathe, globalization has its influence. Too, no field of work or study has been untouched by globalization . If you understand how globalization is shaping your field, you will have an advantage over others. Globalization is worth your attention.”¹⁴

In conclusion globalization brought up new aspects. Technological innovations from the fire and the wheel of early

humans to today's computers have been responsible for the massive change. Accelerating pace of technical progress and development lead social media to immense role on public and direct global political and social process. Thus media has three main function: to inform, to shape public opinion and the commercial function. Also mass media enables the audience to be linked closer to their groups by the sharing of common experiences. What people read in the papers or see something on television or in the cinema, they discuss, analyze and surely the interaction leaves an imprint in their minds.

This enhanced role of media and consequently increased public access to information get the positive appearance and negative as well. On the one hand it means that people have more opportunity to be involved in global processes, be informed about current affaires. On the other hand social technologies can be used by government and interests groups to manipulate and achieve their own goals.

Bibliography

- George T. Duncan, Stephen F. Roehrig, *Reconciling Information Privacy and Information Access in a Globalized Technology Society*, Chapter 5.14, Carnegie Mellon University, USA, 2009
- Howard, Philip N. ,Aiden Duffy, Deen Freelon, Muzammil Hussain, Will Mari, and Marwa Mazaid 2011 *Opening Closed Regimes: What Was the Role of Social Media During the Arab Spring?* Project on Information Technology & Political Islam 2011
- Ismail, Benjamin *Flow of Information and Government Control, North Korea: Frontiers of Censorship*, Investigation Report, October 2011.
- Keohane, Robert O. and Joseph S. Nye, *Power and Interdependence in the information age*, Foreign Affairs vol. 77 No. 5 (Sep-Oct, 1998),
- Karlsen Rune, *Does new media technology drove election campaign change?* Information Polity 15 (2010) 215- 225
- Keohane,Robert O, Joseph S. Nye, *What's new? What's Not? (And So What ?)* Foreign Policy, No. 118 , Spring, 2000
- Lulu, Jack, *Globalization & Media, Global Village of Babel*, Rowman&Littlefield Publishers, USA,2012, p.9
- Spaar,Jonathan, *North Korea: Is Kim Jong-Un “Gangster” Enough to Maintain the Status Quo?* Nk News, April 9, 2012, <http://www.nknews.org/2012/04/north-korea-is-kim-jong-un-gangster-enough-to-maintain-the-status-quo/>
- Munteanu,Nicoleta, *Effects of Globalization on Mass Communication*, Read Periodical, October 1, 2011 <http://www.readperiodicals.com/201110/2561867691.html>
- Li Evans, *Barack Obama The First Social Media President?* Search Marketing Gurus, November 5, 2008 <http://searchmarketinggurus.com/2008/11/barack-obama-the-first-social-media-president.html>
- How Bad Is It for Obama That Oprah Won't Campaign for Him?* The Atlantic April 2, 2012 <http://www.theatlantic.com/politics/archive/2012/04/how-bad-is-it-for-obama-that-oprah-wont-campaign-for-him/255357/>

¹³ Jonathan Spaar, *North Korea: Is Kim Jong-Un “Gangster” Enough to Maintain the Status Quo?* Nk News, April 9, 2012, <http://www.nknews.org/2012/04/north-korea-is-kim-jong-un-gangster-enough-to-maintain-the-status-quo/>

¹⁴ Jack Lulu, *Globalization & Media, Global Village of Babel*, Rowman&Littlefield Publishers, USA,2012, p.9

Cloud computing for business

Khayyam H. MASIYEV

Qafqaz University
Baku, Azerbaijan
xmesiyev@qu.edu.az

Ilkin QASYMOV, Vusale BAKHISHOVA, Mammad BAHRI

Qafqaz University
Baku, Azerbaijan
ilkins.email@gmail.com, baxisova_vusale@mail.ru, mammadb@bakcell.com

Abstract— Nowadays the implementation of information technologies in organizations is a high resource-consuming process associated firstly with the effect of high costs and uncertainty on organizational performance and, secondly, with storage and maintenance issues. This paper will overview the usage of cloud computing in business as one of the effective ways of dealing with the issues mentioned above by describing the essentials of cloud computing, business processes and their convergence. Essentials of cloud computing covers definition, layers, types, advantages and disadvantages of it, while business processes covers business activities, types of business processes and their management, and finally convergence reveals the offerings of cloud computing for businesses, as well as the integration of cloud infrastructure to business processes and practical outcomes from the usage of cloud computing in real-word companies.

Index Terms— cloud computing, business, hardware, software, high costs, infrastructure, computation, performance.

I. INTRODUCTION

The problem of rising costs and uncertainty reinforced by the issues of deployment of technologies requires considerable attention of managers. Lack of rationale on decision making associated with the implementation of information technologies will lead to extreme costs and increased complexity of systems and also, eventually, to a broadening gap between business and technologies – misalignment. This situation points out to the necessity of cost-effective and easily maintained solutions providing considerable benefits for organizations. The idea is to achieve efficiency, streamlining, decentralization, and improved customer relationship management (CRM).

Cloud computing solutions are powerful instruments designed to realize this idea in the form of services. Hardware and software, infrastructure and platform solutions can be offered as a single service accessible via the Internet. Giants like Amazon, Google, Cisco, IBM, Oracle and also the Federal Government of the United States, the Cabinet Office of the United Kingdom have defined their own strategies to implement cloud computing solutions in business, public and government sectors to deliver efficiency, agility, innovation.

A. Abbreviations and Acronyms

- CRM – customer relationship management

- IaaS – infrastructure as a Service
- SaaS – software as a Service
- PaaS – platforms as a Service
- XaaS – everything as a service
- BPR – business process reengineering
- BPM – business process management
- BPMS - business process management system
- TQM – total quality management
- LAP – language actions perspective
- VPN – virtual private network
- PDF – portable document format
- ODF – open document format

II. THE ESSENTIALS OF CLOUD COMPUTING

A. The meaning of cloud computing

Cloud computing can be defined as “an all-inclusive solution in which all computing resources (hardware, software, networking, storage, and so on) are provided rapidly to users as demand dictates”. This definition provides the following specifics of cloud computing:

- All-inclusiveness – multiple solutions are provided in a single one service (hardware, software, infrastructure, operations systems, etc).
- Remote access – users of cloud computing do have access to their data via remote connection
- Rapidity – computational resources are available anytime by request.

With cloud computing users do not require an extremely powerful computer to handle large volumes of data. Instead they can have an internet connection to the server providing the whole infrastructure for handling user data.

Cloud computing has the following key characteristics:

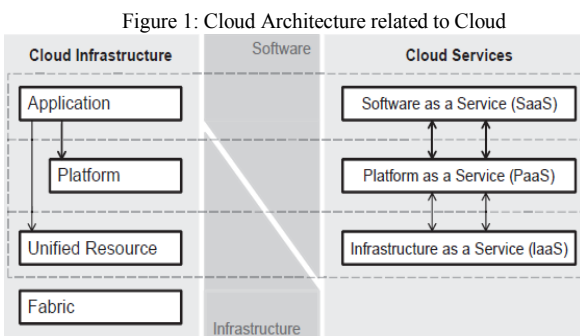
- Computation as a Service – computation is remote. This process is offered as a service.
- Transparency – the structure and the process of data handling is transparent to users.
- Significant cost reduction in hardware and software – users may not need to purchase hardware and software solution to manage data. It is done by remote infrastructure.

- Elasticity and scalability – users can easily purchase computation resources and scale it up to particular extent.
- Disaster-proof and business continuity – remote infrastructures can be secure from force-major occurrences.

The greatest effect of cloud computing in today's organization as mentioned is cost-efficiency. In the long run it also triggers economies of scale for a company. As their business becomes bigger and more complex, the "cloud" will need to become bigger and the system, using economies of scale, will become bigger without even knowing about it.

B. The three layers of cloud computing

Cloud computing solutions are composed of cloud services offered by a service provider. Currently three major services are available:



- Infrastructure as a service (IaaS) – At this level, the product is the hardware and related services (data centers, physical hardware, networking equipment and firewalls, etc). General processing, servers, storage devices, database management, and all other hardware-related services offered as a service to the end user.
- Platform as a service (PaaS) – this layer offers hardware-independent solutions to software developers (operating systems, virtual machines, infrastructure softwares). Developers can write their applications according to the specifications of a particular platform without needing to worry about the underlying hardware infrastructure (IaaS). standardized interfaces and a development platform for the SaaS layer.
- Software as a service (SaaS) – most visible layer of cloud computing for end-users, SaaS, is software that is owned, delivered and managed remotely by one or more providers and that is offered in a pay-per-use manner, because it is about the actual software applications that are accessed and used.

Users often demand not a single cloud computing solution, but all solutions in a single service. It is called XaaS, meaning "everything as a service". Thus XaaS includes IaaS, PaaS and SaaS.

C. Types of clouds

Cloud Computing can be classified and deployed to end customers as Public, Private or Hybrid clouds. Organizations choose deployment models for IT solutions based on their specific business, operational and technical requirements. Let us consider each type:

- Public clouds – public clouds are cloud services provided by third parties but hosted and managed by the service providers. Providers are responsible for installation, management, provisioning and maintenance. Customers access and use the services and physical resources and are charged only for the resources and services they use.
- Private clouds – private clouds are proprietary networks, often data centers for the exclusive use of the organization. These are shared environments built on highly efficient, automated and virtualized infrastructures..
- Hybrid clouds - combination of Private and Public Clouds. They combine on-demand external capacity with on-premises resources and in-house compliance. In this case, the management responsibilities are often split between the enterprise and the public cloud providers, which can often become an issue of concern.

D. Drawbacks of cloud computing

Cloud computing might be seen as a remedial solution to all problems but despite its advantages has several considerable drawbacks that should be taken into consideration:

- Security – issues related to the location of data, its accessibility to third parties and its sustainability to losses.
- Cost – issues related to some hidden costs and future charges for the customers.
- Integration – this problem is related to the obstacles the company may face while integrating cloud computing to the business architecture.
- Knowledge – integration of cloud computing may indeed require special knowledge and skills and even may lead up to higher costs associated with attraction of specialists.
- Flexibility – cloud computing in contrast to private infrastructure may not provide fair feature customization.

III. BUSINESS PROCESSES

A. The meaning of business processes

Business processes are the set of activities that deliver value to the customers. The main features of business processes are following:

- Large and complex, involving the flow of materials, information and business commitments.
- Very flexible, responding to demands from customers and to changing market conditions.

- Long running –It means a single process can take much more time in order to accomplish it.
- Automated – Routine and standard tasks should be performed by computers and other automated systems in order to reduce the time and energy.
- People perform tasks that are too unstructured to delegate to a computer or that require personal interaction with customers.
- Difficult to make visible. In many companies the processes are not conscious or explicit, but undocumented and implicit, embedded in the history of the organization.

Business process management is the capability to discover, design, deploy, execute, interact with, operate, optimize and analyze end to end processes at the level of business design, not technical implementation.

There are two kinds of approach to business processes: Transformation and Coordination approach. According to Transformation approach business processes consist of transformations of inputs to outputs. Coordination approach is also called the Language Action Perspective.

There are several business process methods based on the coordination approach. One of them is Action Workflow which was developed by Medina-Mora in 1992.

TABLE I. ADVANTAGES OF BPM

Advantages of BPM	
Direct	Indirect
1. Editing processes in real time	1. The reduction of the production cycle
2. Reducing internal and overhead expense	2. Improved accuracy of forecasting
3. Automation of key decisions	3. Improving the quality of customer service
4. Reduced maintenance costs	4. Process optimization of supply expenses
5. The decrease in operating expenses	
6. Increase in productivity	

IV. THE ROLE OF CLOUD COMPUTING IN THE BUSINESS

A. The offerings of cloud computing for the business

As mentioned before cloud computing is an effective solution to ever existing problem of complexity of information systems within business domain. Actually if left unmanaged, complexity of technologies rises rapidly leading to an even more misalignment between business and information technologies. In this case cloud computing offers clarification (simplifications) or complexity reduction. By virtualizing the infrastructure, getting resources (either hardware or software) as a remote service, organizations can significantly reduce the level of complexity and count on higher cost cutting. Cloud computing offers a different way to procure and use software and computing services. The continuing evolution of technology will enable cloud computing to make possible a complete breakthrough in the way IT services are provisioned and consumed. Improvements in security, bandwidth, technology standards and virtualization have will motivate to use cloud computing more frequently.

Figure 2: The Action Workflow loop

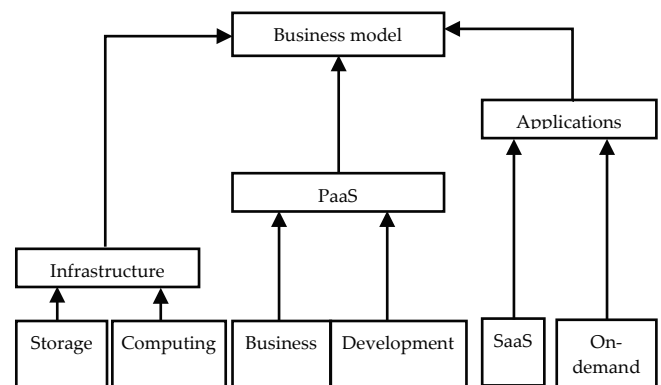


B. Business process management

Business process management (BPM) is a system approach to an organization's processes. Its purpose is to make the business more competitive and successful by better serving its customers. BPM focuses process effectiveness, efficiency and flexibility to adapt to the constantly changing business environment.

BPM is most effective in process intensive industries such as health care, insurance, finance, utilities and government. These businesses rely on human knowledge, information databases and process flows to produce an end result, such as a home loan or a business license. The main advantages of BPM are described following table:

Figure 3: Cloud business model framework



Another offerings of cloud computing are related to responsiveness, connection and specialization. Responsiveness of business comes out of the idea that modern business requires massive data handling and sharing. Thus, high performance hardware and software is required to carry it out.

Organizational resources can be directed at short and long term strategic goals and innovations rather than at secondary problems.

Cloud also makes exploration and entry into new geographic markets and product segments faster, cheaper and less risky, thus, providing secured business.

B. Cloud computing as a business model

A business model refers to ways of creating value for customers, and to the way in which a business turns market opportunities into profit through sets of actors, activities and collaboration. A business model describes how inputs (resources) are transformed into outputs (value). Cloud computing can be considered as an emerging business model, as it provides layers of purchasable services offered to end users. The Cloud computing business model includes three elements according to the layers of cloud computing:

- Infrastructure;
- Platform-as-a-service;
- Applications;

The three elements of the cloud business model framework describe organizations business model according to a type of service. The framework reveals different roles of the business among customers. Organizations can decide on the specific direction(s) of business:

- Infrastructure provider;
- Platform provider;
- Service provider;
- Aggregate services provider;
- Consulting;

In fact all of abovementioned roles can be combined into a single value chain – cloud computing value chain.

C. Integration of cloud computing into business processes

As effective solution to many obstacles in modern business cloud computing can be integrated into business processes.

In practice prominent companies like Amazon, Microsoft, and Google, Dropbox etc. offer cloud computing solution, typically SaaS and PaaS.

Amazon's Cloud Drive is an online storage solution that allows users to upload and manage various types of data files including music, videos, photos, and documents over the internet to store on Amazon's secure servers. Brief features of this service:

Google Drive is another cloud computing solution offered by Google with the first 5 GB of stuff for free, access everything in Google Drive from all devices and file synchronization. 25GB cost \$2.49 and 16 TB cost \$799.99 per month.

Microsoft SkyDrive offers free file storage and access up to 7 GB, with accessing files on the go, simple file sharing and other services. 20 GB of storage cost \$20, and maximum 100 GB cost \$50.

CONCLUSION

Cloud computing is rapidly evolving technology with a considerable value for business. If properly implemented and integrated it can be quite benefitting for businesses. More and more companies offer IaaS, SaaS, PaaS or even XaaS to create business values and to attract customers.

Despite advantages of cloud computing its drawbacks are quite serious and risky to deal with. It is especially important for business operations with considerable investments in them. Security in cloud computing is its major disadvantage. Service providers need to put efforts to secure personal and business data and provide committed level of reliability. Without security concerns cloud computing may not justify business' expectations.

REFERENCES

- [1] Dustin Amrhein, Scott Quint: Cloud computing for the enterprise: Part 1: Capturing the cloud, IBM WebSphere Developer Technical Journal, 2009
- [2] The Art of Service: Cloud computing - the complete cornerstone guide to cloud computing best practices
- [3] Zaigham Mahmood Richard Hill editors: Cloud computing for enterprise architectures, Springer London Dordrecht Heidelberg New York, 2011.
- [4] Katarina Stanoevska Slabeva, Thomas Wozniak, Santi Ristol: Grid and Cloud Computing. A Business Perspective on Technology and Applications, Springer Heidelberg Dordrecht London New York, 2010
- [5] Mikael Lind: Determination of Business Process Types Founded in Transformation and Coordination, International Journal on Communication, Information Technology and Work, Vol. 2 (2006), No. 1, pp. 60–81, page 4.
- [6] CSC's research services: The Emergence of Business Process Management report, January 2002, Version 1.0, page 8.
- [7] Michael McClellan: Business Process Management in a Manufacturing Enterprise, Collaboration Synergies Inc
- [8] Master Thesis: Financial Aspects of Cloud Computing Business Models. Information Systems Science by Jaakko Jäättmäa, 2010.
- [9] Vivek Kundra U.S. Chief Information Officer: Federal Cloud Computing Strategy, The White House, Washington, February 8, 2011.
- [10] Cabinet Office 70 Whitehall, London SW1A 2AS: Government ICT Strategy, Crown copyright, March 2011.
- [11] Victor Chang, David Bacigalupo, Gary Wills, David De Roure: A Categorisation of Cloud Computing Business Models, School of Electronics and Computer Science, University of Southampton, Southampton SO17 1BJ. United Kingdom

Genetic algorithm approach in the minimization of the risk of financial portfolio

P. Dvalishvili, B. Midodashvili
 Department of Computer Sciences
 Iv. Javakhishvili Tbilisi State University
 Tbilisi, Georgia
 e-mail: pridon.dvalishvili@tsu.ge bidzina.midodashvili@tsu.ge

Abstract—In this work we consider the problem of minimization of the risk of financial portfolio. We offer a solution to this problem using genetic algorithm. The program compiled in C++ successfully solves the above mentioned problem for the given input data

Keywords- portfolio of financial assets; risk of portfolio; genetic algorithm; computer program

I. INTRODUCTION

For the appraisal of the portfolio of financial assets there are usually used two criteria, the first is linear – the profitableness of portfolio; the second is quadratic – risk of portfolio [1].

The profitableness of portfolio R_p , when incomes on the risky assets are random variables, can be written as follows:

$$R_p = \sum_{i=1}^N W_i E(Z_i)$$

where W_i is the specific weight of the i -th asset, $E(Z_i)$ is the expected profitableness of the i -th asset.

The riskiness of one asset is measured by the dispersion of income on this asset, and the risk of portfolio – by dispersion of the incomes of portfolio. In order to measure the risk of portfolio, it is necessary to know not only the variation in the incomes of separate securities, but also the degree with which the incomes of pairs of assets change together, i.e., the covariance of the incomes of each pair of assets in the portfolio. Therefore, the risk of portfolio (dispersion of the incomes of portfolio) σ_p^2 , taking into account covariance, is calculated by formula [2]:

$$\sigma_p^2 = \sum_{i=1}^N W_i^2 \sigma_i^2 + \sum_{i=1}^N \sum_{j=1, j \neq i}^N W_i W_j \text{cov}_{ij} \quad , \quad (1)$$

where $\sigma_i^2 = \text{cov}_{ii}$ is the dispersion of i -th asset, cov_{ij} is the covariance between the assets i and j .

Let $\Omega = (\text{cov}_{ij})$, $i, j = 1, \dots, N$ be the dispersion-covariance matrix of assets. Let us denote $W = (w_1, w_2, \dots, w_N)$, W^T be a transposed to W . Then the risk of portfolio (1) can be written in the matrix form:

$$\sigma_p^2 = W \Omega W^T .$$

The problem of optimization of portfolio consists of the following: determine the share of investment according to the assets so that the value of the expected income and the level of risk would optimally correspond to the goals of investor. For example, a goal of investor can be the minimization of risk, but with some minimum income and constraints on shares, which can be invested into each asset.

According to the different goals of investor we receive different kind optimization problems.

Let us consider the case, when we have three assets a , b and c . Investor wants to minimize the risk of portfolio, and at the same time he wants to keep the income level exceeding R . Then the following problem of nonlinear programming is received

$$\begin{cases} Z = W \Omega W^T \rightarrow \min \\ w_a + w_b + w_c = 1 \\ w_a E(z_a) + w_b E(z_b) + w_c E(z_c) \geq R \\ w_a \geq 0, w_b \geq 0, w_c \geq 0. \end{cases}$$

Frequently there are established limits for the investment in the concrete assets, for example, $w_a \leq 0,2$.

For the solution of such problems by the methods of mathematical programming it must be constructed Lagrange's function and found the points of Kuhn-Tucker [3]. Calculation of the derivatives of Lagrangian is also necessary. These are very cumbersome calculations. The analytical solution is not practically possible. Frequently, portfolio problems are solved by the method of Dantzig-Wolfe [2]. There are many powerful algorithms, computer packets, but there is no one method by which it will be possible to avoid the calculations stated above for any similar problem encountered.

For solution of the different problems of optimization, in order to avoid the difficulties stated above, we offer to use genetic algorithm.

Below we consider the specific problem, which is solved by genetic algorithm.

Let us assume that we have three assets a , b and c with expected incomes: 0,11; 0,15; 0,08, respectively. Let dispersion-covariance matrix Ω take the form:

$$\Omega = \begin{bmatrix} 0.00015 & 0.00005 & -0.00007 \\ 0.00005 & 0.00025 & -0.00003 \\ -0.00007 & -0.00003 & 0.0001 \end{bmatrix}.$$

It is necessary to find w_a , w_b , w_c such that to obtain income 0,11, with minimum risk. The problem takes the form:

$$z = 0.00015w_a^2 + 0.00025w_b^2 + 0.0001w_c^2 + 0.0001w_a w_b - 0.00014w_a w_c - 0.00006w_b w_c \rightarrow \min$$

$$\begin{cases} w_a + w_b + w_c = 1 \\ 0.11w_a + 0.15w_b + 0.08w_c = 0.11 \\ w_a \geq 0, w_b \geq 0, w_c \geq 0. \end{cases}$$

The exact solution of this problem is

$$w_a = 0.328, \quad w_b = 0.288, \quad w_c = 0.384.$$

Let us solve this problem by genetic algorithm.

II. GENETIC ALGORITHM

Genetic Algorithm (GA), developed by John Holland [4], is an efficient search and optimization techniques inspired by the Darwin's theory of natural evolution.

The evolution process in the genetic algorithm is done with a population of individuals represented by chromosomes, parameters encoded to the string, bits or other data representation.

Since the first population does not have the final solution, there is a need for keeping an artificial diversity in the population. Diversity can be maintained by using the crossover and mutation operations.

The crossover in the natural evolutionary process means that child will inherit its properties (genes) from its

parents. In genetic algorithms, the crossover operation is needed to mix and inherit good gene combinations from the current population to the new population.

The mutation is performed by applying a random change to the individual's chromosomes. A mutation usually affects only few genes.

Usually the genetic algorithm performs with the following cycle:

1. Evaluate the fitness value for all the individuals in current population.
2. Create new population by using crossover; mutation and reproduction operations.
3. Discard the old population and continue iteration.

III. IMPLEMENTATION OF THE GENETIC ALGORITHM FOR THE PROBLEM OF MINIMIZATION OF THE RISK OF FINANCIAL PORTFOLIO

Genetic Algorithm instead of improving one possible solution works with the population of possible solutions called chromosomes. Chromosomes in our case are represented by binary strings. Since the precision of solutions we require equals 10^{-6} then the chromosomes should contain 21 digits "0" or "1".

From the data of the problem it is clear that we need to have only one binary string to represent a solution of the problem.

The initialization of a population is implemented by randomly generated chromosomes. The fitness of the chromosome is equal to the objective function of the problem.

A crossover operation combines data in the binary strings of two parents, and then it creates a new chromosome. A crossover splits binary strings of both parents in parts of random size. The number of parts in our case equals 2. Then, it alternately copies parts from parents to the new chromosome.

A mutation operation takes a bit in the binary string randomly and replaces it by the 'opposite' bit.

The genetic crossover and mutation operations have the rates 0.8 and 0.05, respectively. We implement also elitism with rate 0.1.

IV CONCLUSION

The algorithm for each generation, consisting of $n=100$ chromosomes, performs the next operations:

1. Selects $n/10$ best chromosomes of the population.
2. Randomly selects pairs of parents from the best half of the current population, produces new chromosomes by performing a crossover operation on the pair of parents and a mutation operation on new chromosomes, and adds them to the elite chromosomes in population.

The algorithm is repeated until finite iteration number is reached.

The algorithm described above has been applied to the program written in C++ and successfully tested with different parameter values. Genetic algorithm appears to find a good solution for the problem of nonlinear programming under consideration, and the convergence of the algorithm depends on the way the problem is encoded and which crossover and mutation methods are used.

REFERENCES

- [1] Markowitz H.M. Portfolio Selection. *Journal of Finance*, 7, 77-91, 1952.
- [2] Terry J. Watsham and Keith Parramore. *Quantitative Methods In finance*. International Thomson Business Press. 1999.
- [3] Fletcher R. *Practical Methods of Optimization*, 2nd edn. John Wiley, New-York. 1987.
- [4] Holland, J., *Adaptation in natural and artificial systems*, 1975. University of Michigan Press, Ann Arbor.

Intellectual Support System of EIA (Environmental Impact Assessment) Procedure in Region of Caspian Sea

R.A. Karayev¹, K.A. Aliyev², M.A. Nagiev³, N.E. Kazimova³

¹Institute of Cybernetics NASA, ²Oceaneering International Services LTD, ³“USTAD” LLC
Baku, Azerbaijan

karayevr@rambler.ru, kaliyev@oceaneering.com, miragabey@yahoo.com, kazimova_n@yahoo.com

Abstract – Are considered the opportunity of modern methods usage of knowledge engineering, of ecological monitoring, as well as ICT for solution of one of the most actual problems in the Caspian Sea region – problem of “ Environmental Impact Assessment in transboundary context”, which has become critically important by collapse of USSR.

Keywords – Caspian Sea, transboundary Environmental Impact Assessment, intellectual support system

I. INTRODUCTION

Environment Impact Assessment (EIA) is a national assessment procedure of possible impact of planned activity on environment. According to the recommendations of SCOPE (Scientific Committee on Problems of the Environment) the EIA procedure is a legally affirmed element of economic planning in many countries of the world as well as in the Caspian Sea region countries.

The planned activity (PA) is considered as any activity, which requires decision-making by competent organs according to the applied national EIA procedure.

Transboundary impact means any impact in the region, which is under jurisdiction of one country that realises the planned activity in the territory of another country [1, 2].

Nowadays there are lots of projects with possible transboundary impact on environment of Caspian Sea region. (projects of oil and gas extraction, projects of Caspian Sea region countries, which are connected with rivers that flow into the sea). Therefore, the necessity of common normative-methodical documents on environment impact assessment in transboundary context (EIAT) is occurred. Absence of such documents cause the problems not only in protection of natural ecosystem of Caspian Sea region, but also for the workers-out of the projects, who want to be sure in observance of all legislation requirements as well as in national and international relations field.

At the moment there are four main documents concerning the matters [1, 2, 3, 4]. In these documents as in other documents on EIA, are normalized organization interaction regulations, but they do not contain the scientific-methodical support instruments of EIAT procedure. The methods of EIA, which are suggested in [4] and [5], show the

condition of ecology science of 30 years-old. Methods, which are used in projects of oil-gas consortiums [6, 7], are based on EIA paradigms that were turned out in the open water bodies (Gulf of Mexico, Persian Gulf, North Sea and etc.) and solve local issues concerning EIA concrete contract areas. EIAT procedure pursues other objectives and solves another more complicated issue. For many regions of Caspian Sea it is aggravated by two circumstances. Firstly, by wide range of anthropogenic and natural factors of impact (fig. 1). Secondly, by increasing interaction of natural and anthropogenic sources of pollution by reduction of self-cleaning ability and ecological capacity of the sea.



Fig.1. Anthropogenic and natural factors of impact on Caspian Sea ecosystem. Source: www.grida.no/res/site/file/publications/vitalcaspian.pdf

Absence of the scientific-methodical support instruments of EIAT procedure, orientation on ecological practice of open water bodies leave the wide space for subjective interpretation as well as for possible misunderstandings in interethnic ecological conflicts.

There is a necessity of new adequate approaches and new support instruments of EIAT procedure. Below is mentioned common arrangement of EIAT task and description of intellectual support system of EIAT, which can be realised on the basis of modern methods of knowledge engineering, ecological monitoring and ICT means.

II. EIAT TASK ARRANGEMENT

EIAT task arrangement considers, firstly, specific conditions of Caspian, where along with impacts of planned activity, there is impacts of natural and anthropogenic sources of pollution and, secondly, considers whole life cycle of planned projects. So, for example, in case of offshore projects ecological consequences of cancellation and liquidation work of borehole equipments, water development facilities, pipeline transportations. In general, EIAT task arrangement can be presented as following:

$$\Delta_0(I, S) \xrightarrow{D(P, R, t_0, t_n)} \Delta_n(I, S),$$

where $\Delta_0(I, S)$ – assessment of initial condition of environment of the affected district;
 $\Delta_n(I, S)$ – forecast of expected impact of planned activity on environment of the affected district;
 P – information about the project of planned activity of the country and expected transboundary impacts on environment of affected district;
 R – information about sources of pollution;
 I – environment impact objects and indicators;
 S – environment impact assessment scales;
 $D(P, R, t_0, t_n)$ – EIAT operator, which realizes assessment of transboundary impact character and scale;
 t_0, t_n – starting and ending time of planned activity project.

III. SHORT CHARACTERISTICS OF APPROACH

Till now there is no common theory of EIA. In their practical activity the drafters of Caspian projects orient themselves to the “best world practice”, which uses

- structured templates (frames), which were approved in marine oil and gas production for logistic analysis of possible impacts in normal, non-staff and emergency situations,
- “assessment indexes/indicators system” of

environment condition (of water, soil, ground accumulation, flora and fauna, air) [8, 9],

- expertise assessments of possible impacts on separate indicators,
- modelling methods of oil overflow, emissions into atmosphere and drilling sludge, deposit water, chemical reagents escape into marine environment,
- traditional procedures of compression of expert judgements compression (collective discussions by involving public, numerical scores, “method of weighted total” and etc.)

It is not enough to use these instruments for solution of EIAT issues in Caspian Sea condition. Condition indexes/indicators and results of modelling assess separate elements of environment, but they give the common whole map which is necessary for long-term analytical predictions and decision-making. Efforts for transferring to common assessments by involving various models of compression cause “the loss” of the information as well as subjective interpretation that is connected with personal, group or corporate interests.

Sufficient step in solution of issue can be caused by creating of intellectual support system. The basis of system consists of:

1) ecosystem paradigm of EIAT [10, 11], which considers

- phenomenological characteristics of water ecosystem [12],
- Tehran Framework Convention on Environment Conservation of Caspian Sea (2003),
- specific character of the Caspian Sea – closed basin, sea level fluctuations, durational historical pollution, reducing ability of self-clearing, intense condition of main bio-resources, in the first place unique sturgeons, transboundary character of stocker and seasonal migration of sturgeons, large-scaled poaching,
- leading practice of ecotoxicological normalization of marine operations (in HOCNF format (CEFAS, United Kingdom) [6]),

2) modern methods of knowledge engineering and ICT means, which give unique opportunity,

- to involve almost whole ecological practice of marine oil and gas production (international and regional) for solution of the issue,
- to increase accuracy of EIAT by integrated usage of problem knowledge and to create “cyber consultants” on this basis, whose competency can be sufficiently higher than the competency of separate experts and expert groups.

Demonstrative prototype of support system, which is based on such an integrated approach, is fulfilled as a “system cover” of frame type and as a situational qualimetry mechanisms offered by prof. D.A. Pospelov. We have worked out the applied version of this mechanism in cooperation with “GIPROMORNEFTEQAZ” (SOCAR, Azerbaijan)

and Amoco Company (USA). These instruments allow to creating situational models of PD project, environment of affected district, to create clustery of expected ecological consequences and to fulfil quality predictions of EIAT on various sceneries in “situation network”.

The prediction searches in “situation networks” can be uncompleted. In this case system recommends application to “joint” procedures: “imitative modelling of oil overflow”, “imitative modelling of drilling sludge escape”, “imitative modelling of emissions into the atmosphere”, “cognitive modelling of long-term dynamics of transboundary impact” and etc.

Total assessment of transboundary impact is built up according to “evidence and scientific predictions sums, which let us assess the danger of planned economical activity” and, it includes the following three:

$$EIAT: = \langle S, I, C \rangle,$$

where S – linguistics scale of transboundary impact (ZERO, WEAK, MILD, STRONG, CATASTROPHIC);

I – complex of evidences and scientific predictions of impact which are specific for affected district (at the moment such complexes are built up for various water areas of Azerbaijan sector – FORESHORE (five zones), SHELF (three zones), RECESSION OF DEPTHS);

C – coefficient of confidence in assessment calculated according to “evidence consolidation law” of D. Shortleaf.

Unlikely the methods accepted in international practice the EIA system allows to take into account structure-functional organization of concrete biogeocenosis, the regularity of their restructuring, their steadiness scale and their apprehensibility to anthropogenic load. It can be achieved by setting the “system cover” on the conditions of the affected district and on the parameters of PD project. The setting is realized by question-answer interface, which includes “situational map” of EIAT issue and Web-manual, that allows to acquire necessary data from Internet.

The presented approach creates methodological premises for holding unified ecological politics in Caspian Sea region as well as for creating common method basis of EIAT, which can be recommended to project developers and competent authorities of Caspian Sea countries.

IV. INTELLECTUAL SUPPORT SYSTEM OF EIAT PROCEDURE

System realizes the operator D and corresponds soft- and hardware complex, which has connections with local and regional ecological monitoring system (REM) of Caspian basin, with nature-conservation organs of Caspian countries, scientific-research and

project institutes, hydrometeorological survey services, and space remote sensing system of the Earth, ecological departments of oil companies. On the fig. 2 is illustrated the PC-oriented structure of main frame of system built up for Azerbaijan sector of the sea.

V. SYSTEM APPLIED OPPORTUNITIES:

1. Ecological security provision of national sectors of the sea by working out of scientific basis and improvement of EIAT methods,
2. Scientific-methodical support provision by arbitration examination of conflicts between “the country of origin” and “affected country”, as well as in case of interethnic ecological conflicts, which are inevitable in the condition of large-scale extraction and of increasing geopolitical intensity in the region,
3. Provision of objective and timely information about possible transboundary impacts for taking preventive and efficient reaction measures,
4. Consolidation of international cooperation in the EIAT field in working out of economic plans, social-economic development and nature-conservative legislation programmes in Caspian Sea region.

VI. CONCLUSION

Intense ecological circumstance in Caspian Sea region and scales of anthropogenic load shows that researches of the EIAT problem can be considered as priorities of regional science. They will determine the efficiency of ecological catastrophes prevention efforts in Caspian Sea region. These researches completely meet the requirements of Espoo Convention [1, p. 39]: “to place special emphasis to working out or more active implementation of concrete research programmes aimed at improvement of existing quality and quantity methods of consequence assessment of planned activity”.

EIAT procedure support instruments can be used in calculation of compensatory payments among neighbour countries, from which territory transboundary transportation of pollutants is realized. Developments of such instruments is actual and for the reason that regional ecological safety issue is becoming one of the crucial components of geopolitics, when economically developed countries try to manage the hydrocarbon resources of Caspian governments (Russia, Kazakhstan, Azerbaijan, and Turkmenistan).

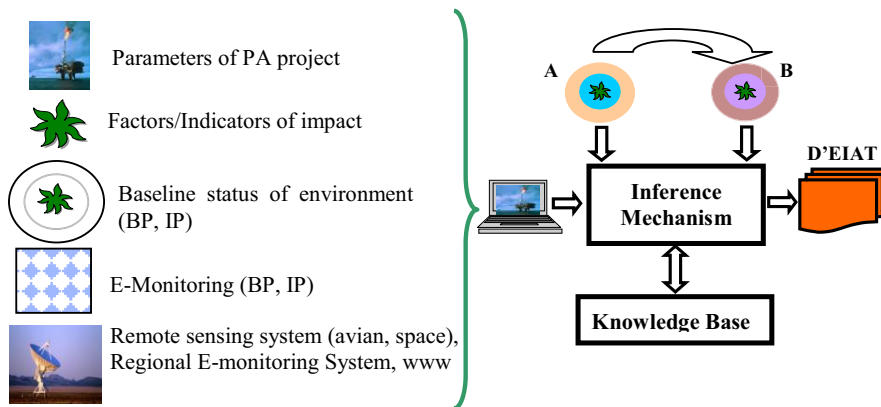


Fig.2. Structure of main frame of EIAT system (South Caspian, Azerbaijan sector)

Notation conventions: *A* – reset state of environment of affected district, *B* – expected state of environment of affected district, *BP* – background polygon, *IP* – impact polygon. Knowledge basis: ecological-landscape map of South Caspian water area, ecological passports of potential sources of pollution and classificatory of toxics in HOCNF format, virtual archive of EIA projects and Caspian Ecological Programme reports (www.caspianenvironment.org), normative-legal and methodological basis of EIAT procedures, contact information.

REFERENCES

- [1] Convention on Environmental Impact Assessment in a Transboundary Context. UN. Espoo. Finland. 1991.
- [2] Environmental Impact Assessment in a Transboundary Context in the Region of Caspian Sea. Guide. UNEP. EBRD, CEP. 2003.
- [3] Protocol on Strategic Environmental Assessment to the Convention on Environmental Impact Assessment in a Transboundary Context. ECE. UN. 2003.
- [4] A guide to Environmental Impact Assessment for the state-participants CIS. UNEP, Center of Intern. Projects. Moscow, 2003.
- [5] A guide to Impact Assessment of planned activity on Environment of Caspian Sea in a transboundary context for the Russian Federation and other Near-Caspian countries-participants. CIS, UNEP. 2003.
- [6] Mater. of the I and II Inter. Scientific and Practical Conference “Problems of the Caspian Ecosystem Conservation Under Conditions of Oil and Gas Fields Development” Moscow-Astrakhan: KaspNIRKH Publ. 2005, 2007.
- [7] GIWA (Global International Water Assessment) Caspian Sea. GIWA Regional Assessment. 23. University of Calmar, 2003.
- [8] Ecological Indicators / McKenzie D.H., Hyatt D.E.(eds.) // Proc. of the International Symposium on Ecological Indicators. Elsevier Appl. Sciences. N. Y. 1990. 1600 p.
- [9] Ott W.R. Environmental Indices: theory and practice // Ann. Arbor Sci. Mich, 2008. 459 p.
- [10] Karayev R.A. IEA the Caspian oil and gas projects: old problems and new decisions // Mater. of the First International Scientific and Practical Conference “Problems of the Caspian Ecosystem Conservation Under Conditions of Oil and Gas Fields Development” Astrakhan: KaspNIRKH Publ. 2005. P. 88-93.
- [11] Karayev R.A. et al. Environmental Monitoring of Caspian Oilfields: New Paradigm. New Solutions // Interdisciplinary Environmental Review. New York: Rowman Publish., V. 6, Issue 1. 2005. P. 71-81.
- [12] Holling C.S. Adaptive Environmental. Assessment and Management. Ed. by C.S.Holling. New York: John Wiley&Sons. 1978. 397 c.

SESSION 4

ICT in Education, Research and Science

AICT 2012

Open-Access Journals (Free Valuable Information on the Web)

Aref Riahi, MA in Information Science,
University of Tehran, Iran (Ariahi@ut.ac.ir),
Call No. 0981525264700

Samaneh Khakmardan, MA in Publishing
Management, University of Imam Reza,
Mashhad, Iran

Abstract - Journals, due to variety and rapid distribution of information, in comparison with other information litters have more addresses. But nowadays, emergence of new Information Technology (IT) such as internet and web has increased the interest and popularity of journals, especially open-access journals. Therefore, the present research intends to investigate open-access movement having been the cause of open-access journals emergence; furthermore, it intends to introduce open-access methods and present these journals' characteristics and definitions. The present research has been done using literature review and library resources, whether traditional or digital. Findings have shown that open-access journals created in response to periodicals provide a good opportunity for libraries and researches to have access to information and researches easily and without paying any money, in this way they can accelerate research process. Free electronic publications provide opportunities for librarians; furthermore, their responsibilities necessitate using this opportunity, and through gathering, organization, and these periodicals' targeted dissemination, as well as frugality at budget, take an effective step in expanding and enriching their libraries.

Keywords: Open access, Journals, Valuable information, Free information,

Introduction

Knowledge of societies' development and progress is the first principle for research implementation. New knowledge is attained through studying

existent and past knowledge, ideas exchange, and establishing interdisciplinary scientific networks. This research purpose is access to research findings for all people and establishment of appropriate footstone for future researches; furthermore, obstacles to attaining scientific results bring about underdevelopment. Therefore, publication of research findings and providing access to these publications is the prerequisite for efficiency of the research process.

In general, access to research findings through subscription to printed or electronic publications is the dominant trend in scientific communication. Universities and research institutes spend an expensive cost for gaining direct access to researches. Furthermore, collection of necessary publications for scientific community has gotten a little more difficult due to; on one hand universities libraries' and researches' budget decrease, on the other hand publications' price increase. Its only outcome is inability of libraries in satisfying scientific board members' and researchers' information needs; undoubtedly it will affect research and development]1[.

Nowadays, researches publish their publications in the form of essays in order to provide easy and rapid access for others. Therefore, scientific publications have a special place in responding to researchers' needs. Journals value, during the time of publication and even after some years, can't be denied. We are more and more observing more relationships among different communities since advent of internet and its various facilities. One of these facilities causing acceleration of human beings communications and ideas exchange is the international network of web. Since 1990,

following web invention, whole text of published essays will be available for researchers electronically and gratuitously in scientific journals. Nowadays, web is one of the major scientific communication channels among researchers. In fact, internet and web have transformed traditional system of scientific communication and the way to produce and distribute information; furthermore, they have changed access pattern to scientific sources and research findings results from the system based on subscription right to the system based on free access. Free access means continuous access to scientific works' whole text without necessity of paying any money to publisher or author, but observing author's moral rights. This research intends not only to investigate open-access journals as information litters, but also indicates the undeniable importance of open-access journals in scientific communications]2[.

In recent years, open-access journals have rapidly transformed into a tool for research communication and a place for publication of scientific texts. Nowadays, a large percent of scientific texts is accessible just in the form of open-access journals. In a research report published by Hess et al in which ideas of 688 publishers on works' publications in the form of open-access are inquired, the general attitudes towards open-access principle is positive. Furthermore, in recent years, the increase in costs allocated to subscription to scientific journals has been one of the obstacles in the way of having access to scientific results. Therefore, open-access pattern to information and publication of free scientific journals for fighting against permanent increase in subscription costs has attracted university and research communities throughout the world. This trend brings benefits for developing countries for which access to scientific journals is difficult (economically), therefore they can compensate lack of access to information to some extent]3[.

Open-access movement: open-access movement to sources started since 1990, gradually developed dispersedly, with different attitudes, in different fields of study, and different places. But the term

open-access was proposed in 2002 in a conference on open-access initiative in Budapest; in addition, whole emphasis was on open-community institute in order to solve the problems of research communication. One of the results on this conference was providing a definition for open-access.

They have said in Budapest conference: "open access to research and scientific texts means these texts' free accessibility through internet for all and each user is allowed to read, transfer to PC, copy, distribute, print, search, and connect to these essays' whole text and use them; furthermore, this is easy without any financial, legal, and technical limitation, other than limitations to access internet itself. The role that here we can mention for author right is authors' control over the integrity of his work, correct identification of that work, and reference to that." Therefore, two approaches are suggested to have open access to research journals:

1) Self-archiving: Steven Harvard mentions self-archiving as a green path for having open access. When authors provide access to their own essays gratuitously in the form of digital and through internet, it is called self-archiving. These essays can include those publications before or after printing.

2) Open-access journals: second approach for open access is open-access journals which are recognized at Budapest conference. Steven Harnad introduced open-access journals as a golden path for open access. These journals have these characteristics: they are research (stories, popular journals, self-study books, and materials like these are not in this category); they are accessible gratuitously; they are digital; access to these are easy through internet; they use quality-control mechanism such as popular journals; they give the authors the possibility of preserving the copyright, but not paying these authors' essays their fees]4[.

Open-access journals: there are many different definitions for open-access journals presented by conferences, but the essence of these are the same. Here two important samples presented by open-access publication experts are mentioned: Charles and Bailey believe open-access journals are electronic journals published by a person or a

company's financial support; readers do not pay any money for using those and they can benefit of services such as studying, downloading, transferring to others, or materials printing. Furthermore, essay acceptance depends on some referees' acceptance, and author remains the real owner of his work's right. Suber defines open-access journals in this way: e-journals are gratuitous, permanent, and free from any legal limitation arise from intellectual property and author right. These journals are published by some non-profit institutes such as "PLoS" or profit institution such as "BioMed Central". Furthermore, Directory of Open-Access Journals (DOAJ) knows open-access journals as a journal using a pattern for distribution in which a reader should pay no cost]5[.

Methods of Funding Open-Access Costs: We can say that there have been different ways suggested and tested for open-access journals' funding, some of the most important ones are: 1) paying the cost by the author: this method that is among the most common ones, authors pay some money to journal for essay publication. May be, at first glance, author's paying money seems strange, specially due to this fact that people expect some money to be paid to the author, like author of a book, for copyright, not paying money for publication of his work. But supporters of this approach believe that these are the authors seeking some ways for introducing their scientific achievements to others, scientific validation, and maintaining their works for future generations. 2) Intercalating advertisements: advertisement intercalation is more effective for journals which have very special readers. For example, we have always been observing medical equipments' advertisements in this field's scientific journals. Advertisements in these fields can have been a good source for open-access journals' publication. 3) Funding by an organization: most open-access journals are dependent on an organization or scientific association and the costs needed for their publications are provided by this institute or association. 4) Organizational membership: in this pattern, an institute usually responsible for several

scientific journals receives membership right through open access, subsequently provide special services for them. Of course, access to these journals is not limited to just a special member organization, ever one throughout the world can have access gratuitously to these journals]6[.

Open-access journals benefits: nowadays, many researchers have found that open-access journals have more benefits than printed journals, the most important ones are: reduction of printing cost in comparison with paper journals and being more economic _ access to update information in all academic and research fields _ increasing access to scientific journals, increasing the number of readers, subsequently increasing number of citations and increasing their impact factors _ observing open-access principle for all readers _ no limitation for print circulation _ free access to open-access journals for researchers and readers _ coordination with the latest developments in the publishing world _ high speed of essay publication in open-access magazines _ increasing the richness of library periodicals.

Author's copyright in open-access journals: open-access journals are founded considering the supposition of observing author's moral right by readers, and publication of an Internet-Russian journals does not necessarily mean lack of author's copyright. Observing moral rights encourages and stimulates authors, subsequently science and knowledge development. Moral right includes author's right for opposing any distortion, transformation, mutilation or adaptation of the work; they may disturb its honor or reputation. Moral right means that society always knows the author as the real inventor of that work; furthermore, it mentions him with this name, and believes that he deserves this rank. Therefore, the most important principles for open-access journals usage are citation principles' observance and resource citations which are author's obvious right. But author's financial right is not mentioned in open-access system and open-access journals do not have subscription right. Here, we should mention that subscription right has always been one of the obstacles in the way of access to scientific

researches results. Therefore, open-access movement, subsequently, open-access journals have been formed.

Libraries, librarians, and open-access journals: librarians and research journals have always been linked to each other deeply and these sources have been identified as components of research or scientific communications system. Research libraries role has been the formation of a series of relevant journals to the clients' needs. In general, libraries have the same duties towards open-access journals as they do for other cases; but librarians may face some special issues because open-access journals' nature is different from other issues to some extent. Here, librarians face acquisition/collection, organization, archive, conservation and preservation, at last distribution processes. In technical words, here collection and acquisition are very commonplace and easy. Libraries can copy data very easily from one place to another. Organization is a little more difficult, but traditional cataloging can have an effective role here. Issues relating to archive, conservation, and preservation have not been solved yet. LOCKSS principle or Lots of Copies Keep Stuff Safe presents some solution, but transferring data from one format to another will be a long-term solution. Distribution is the most difficult issue. Appropriate indexing techniques are very useful here, and inappropriate indexing will not bring about desirable searching. Libraries may provide metadata for open-access texts. But distribution won't be finished by indexing and searching. Richness of data existed in web has risen individuals' expectations. What people need are methods to manage their data and use for attaining their goals. Furthermore, librarians should pay attention to five points as ones cooperating with researchers, open-access journals publishers, and managers in research communications process: 1) open-access journals and texts collection 2) organization 3) archive 4) conservation and preservation 5) providing access to those. We should always mention that these issues provide many opportunities for librarians and librarians should enjoy those]7[.

Conclusion: publications with open access having come into existence in response to periodicals, on one hand, provide a good opportunity for librarians and researches to gain an easy access to scientific and research data gratuitously in the form of free electronic journals, therefore accelerate research process. On the other hand, more visibility of these publications facilitated through world network raises impact factor and causes more citation to those; these all enrich scientific and research communications. Gratuitous E-Journals provide opportunities for librarians and their responsibilities necessitate using those and take effective steps on the way of improving and enriching their libraries through acquisition, organization, and targeted distribution of those publications. Therefore, countries and scientific communities should take effective and fast steps for releasing productions and research and scientific data and try to remove all obstacles with their actions. Furthermore, open-access publications development and these resources' integration in online databases, observing the activity relating to quality control at journals admission, on one hand increases scientific community confidence on the kind of publication; on the other hand, facilitates these resources' location and retrieval during next stages. Besides, we can conclude that open access is necessary for spreading Iranian society's enlightenment, access to knowledge and data, and economic and cultural development, and government should take important steps for expansion and development of open access. On the other hand, readers should encourage authors to publish their essays in open-access journals through observing author's moral law and citation principles.

References:

- [1]. Anderson, Paul. 2007. What is web 2.0? Ideas Technologies and implication for education. JISC Technology and Standard Watch 1 (3) : 1-64.
- [2]. Hegman, M. 2006. Open Access to Scholarly Publication. Retrieved April 6, 2012, from: http://www.soros.org.openaccess/pdf/Melisaa_Hagemann.pdf
- [3]. Hess, T., et al. 2007. Open access & science publishing. Retrived April 6, 2011, from : <http://>

www.wim.bwl.uni-muenchen.de/download_free/sonstiges/mreport_2011_01_pdf

[4]. Harnad, S. 2005. Fast-forward on the green road on open access. The case against mixing up Green and Gold. No.42. Retrieved December 25, 2012, from :
<http://www.ariadne.ac.uk/issu42/harnad/>

[5]. Suber, P. 2005. Bethesda statement on open access publishing, Retrieved April 5, 2012, from:
<http://www.earlham.edu/peters/fos/bethesda.htm>

[6]. Zhang, S.L. 2007. The flavors of open access. OCLC system and services: International digital library perspective 23 (3): 229-234.

[7].Morgan, E.L. 2004. Open access publishing. Retrieved April 6, 2011, from :
http://www.soros.org/openaccess/pdf/open_access_publishing_and_scholarly_societies.pdf

A Key Component Extraction Method Based on HMM and Dependency Parsing

Jianchu Kang¹, Songsong Pang¹, Jian Dong¹, Bowen Du¹, Jian Huang²

State Key Laboratory of Software Development Environment

Beihang University

Beijing, China

¹{kang, pangsongsong, dongjian, du_bowen}@nlsde.buaa.edu.cn

²{huangjian}@buaa.edu.cn

Abstract—Increasing attention has been paid for POI (Point of Interest) data query for travel information service. The correct extraction of key components in question is crucial for improving the accuracy of query results. The paper proposes a key component extraction method based on HMM (Hidden Markov Model) and dependency parsing. Firstly, the sentence pattern classifier is established by HMM. And then, questions are classified by classifier. Finally, combination of sentence pattern's structure, the four key components are extracted by dependency parsing. The results show that the F1-Measure is 0.83, which well proves the effectiveness of the method.

Keywords—HMM; dependency parsing; segmentation; sentence pattern; key component

I. INTRODUCTION

With the arrival of the 3G era, simple traffic information, control information, weather information, etc., are insufficient to meet people's needs. More intelligent personal services are expected, among which POI data query for travel information service is a focus of attention.

Although some search engines do provide services of POI data query for people, lots of query questions cannot be correctly identified, due to their relatively simple identification rules. More key components need to be identified for more accurate query results. The extraction of key components in natural language query questions, which is a tricky issue in the parsing of natural language processing(NLP), is hence critical [1].

In this paper, the key component extraction for query questions about transport travel information faces three main problems which are component identification (segmentation belongs to which of the four components mentioned below), segmentation merger (for example, '北京', '航空', '航天' and '大学' are merged into '北京航空航天大学') and identifying multiple parts of speech of the same component (that segmentation belongs to which of the four components should not be affected by its part of speech).

A method to be used for solving these problems above is parsing. There are two kind of widely used grammatical systems in parsing which are phrase structure grammar and dependency grammar. The paper uses dependency parsing technology to deal with questions because of the following two

reasons. On one hand, natural language queries questions (for example, "北京航空航天大学附近有好吃又便宜的川菜馆吗?") for travel information service have its own characteristics. They generally consist of 5 parts: the geographical name or benchmark name (NS), the spatial relationship vocabulary (ND), the functional vocabulary (FUNC), the POI's name, category or attribute vocabulary (TYPE), the particle or tone composition (TONE). It can be seen that the grammatical structure of such sentences is relatively simple, and that the dependency relationships between the components are obvious. On the other hand, representation of dependency grammar that focuses on reflect the semantic relationship more inclines to human linguistic intuition [2], which is conducive to some upper applications, such as semantic role labeling, information extraction, etc.

However, as questions with different kinds of sentence pattern may have the same syntactic analysis result, component identification and segmentation merger must be done within one sentence pattern. Therefore, questions to be analyzed must be divided into the corresponding sentence pattern to ensure that the third issue mentioned above is resolved, before identifying sentences' components. Question contains different key components and the same component may show different vocabulary, which is fit for the characteristics of HMM.

Therefore, to resolve the three problems mentioned above, the paper firstly uses HMM to classify these questions. And then, these key components are extracted by dependency parsing within the sentence pattern. In the following, section II gives the technique background; section III presents the method of extracting key components; in section IV, some experiment results are showed, which demonstrates the validity of the method; and section V provides the conclusions and future works.

II. TECHNIQUE BACKGROUND

In order to design an effective method of key component extraction, the following two fundamental techniques are involved, that are dependency parsing and HMM. In this section, a brief introduction to the two techniques will be presented.

A. Dependency Parsing

Dependency parsing takes advantage of the parsing tree to save these dependency relationships between segmentations. In this paper, firstly, natural query question's dependency relationships are saved in the tree structure by dependency parsing; secondly, ordered pairs of dependency relationships of the modifier and the modified segmentations are got, according to relationships between parent nodes and child nodes [2,3]; thirdly, combination of ordered pairs, segmentations are mapped to (NS, ND, FUNC, TPYE), which can roughly solve component identification and segmentation merger.

In order to get the dependency relationship of Chinese question, domestic and foreign scholars have done lots of research. At present, more mature approach have unsupervised learning method [4,5], SVM-based shift-reduction method [6], Divide and conquer strategy [7], Corpus-based statistical learning method, etc. In this paper, Ma Jinshan's method will be used [2], that has been integrated into the HIT's language technology platform—LTP system, so the LTP system is used for getting the dependency relationships of query questions.

B. HMM

Through the analysis of section I, the paper needs to solve the classification of natural language query questions. For the classification of Chinese text, more mature algorithms have Naive Bayes [8], SVM(Support Vector Machine) [9], k-nearest neighbor [8], decision trees [10], artificial neural networks [11], HMM [12], etc. The paper chooses HMM to sentence classification.

The following will be a brief introduction to some concepts used in this paper [13].

The set of hidden states is $S = \{s_1, s_2, \dots, s_N\}$. The state in time t is q_t , $q_t \in S$, and the hidden state sequence with the length T is $Q = (q_1, q_2, \dots, q_T)$.

The set of visible symbols is $V = \{v_1, v_2, \dots, v_M\}$, where M is the number visible symbols each state may display. The symbol in time t is o_t , $o_t \in V$, and the observation sequence corresponding to Q is $O = (o_1, o_2, \dots, o_T)$.

Transition probability matrix of the hidden state is $A = (a_{ij})_{NN}$, where a_{ij} is the transition probability from s_i to s_j , as follow:

$$a_{ij} = P(q_{t+1} = s_j | q_t = s_i), 1 \leq i, j \leq N \quad (1)$$

Transition probability matrix from the hidden state to the visible symbol is $B = (b_j(k))_{NM}$, where $b_j(k)$ is the transition probability from s_j to v_k , as follow:

$$b_j(k) = P(o_t = v_k | q_t = s_j), 1 \leq j \leq N, 1 \leq k \leq M \quad (2)$$

A and B are both the characteristic matrices of sentence pattern.

The initial state probability vector is $\pi = \{\pi_1, \pi_2, \dots, \pi_N\}$, where π_i is the initial probability of s_i , as follow:

$$\pi_i = P(q_1 = s_i), 1 \leq i \leq N \quad (3)$$

A HMM can be expressed with $\lambda = (\pi, A, B)$.

HMM have three core issues [14]. The first one is valuation issue predicting question's sentence pattern belongs to, which aims to calculate the probability of generating the specific observation sequence O by a HMM with a_{ij} and $b_j(k)$ known. The second is decoding issue, whose purpose is to determine the most possible Q generating the given O with a HMM known. The third is learning issue extracting sentence pattern's characteristic belongs to, which aims to calculate these values of a_{ij} and $b_j(k)$ by training the real questions collected, based on the premise that general structure of a HMM is known (for example, the number of hidden state and visible symbol are known).

III. THE KEY COMPONENT EXTRACTION METHOD

This section describes the specific process of key component extraction. Fig. 1 illustrates the procedure. The model takes natural language query question as input. To begin with, segmentation sequence and abstract information (POS(Part of speech) tagging sequence corresponding to segmentation sequence and dependency relationships sequence) are got by Chinese segmentation technique and dependency parsing. Then sentence pat-tern classifier is established by training these POS tagging sequences with HMM, and then corresponding relations between segmentations and key components of question can be determined. Finally, the original segmentation results are merged or restructured into (NS, ND, FUNC, TPYE), according to component identification module and segmentation merger module.

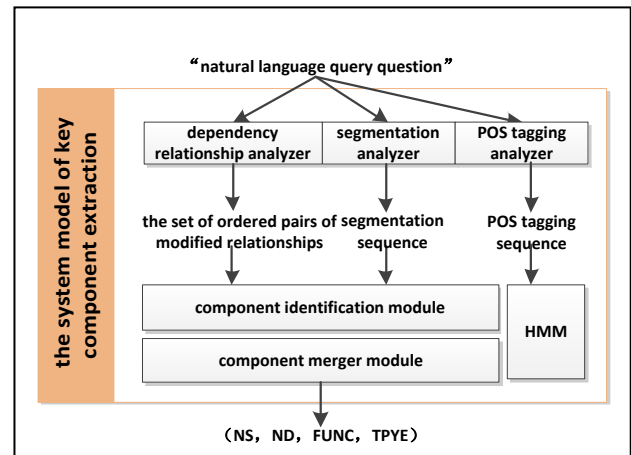


Figure 1. The frame diagram of the system model of key component extraction

This section includes the following five parts of work. The first one is data pretreatment; the second part is to obtain segmentation sequence and abstract information; the third one is to establish the sentence pattern classifier by HMM, which provides the basis for the classification of questions of the fourth part; the fourth part is to classify questions to be tested by classifier; the fifth part is to extract the four kinds of key components of questions, based on sentence pattern's structure and dependency relationship sequence.

A. Data Pretreatment

Although majority of questions contain tone composition, the tone composition has little value in human-computer

interaction, so the tone composition is filtered out in this paper. In order to extract characteristics of sentence patterns, the potential question patterns must be got so that training questions are preliminarily classified. This section will address the two issues.

The paper filters out the tone composition in manual way.

In order to obtain potential sentence patterns, first of all, grammar components are arranged, according to their possible location in question; secondly, we make crossover operation to these components to get all possible and meaningful phrases; finally, we increase overlapping relationships among components by the binomial combination way, so as to get all meaningful sentence patterns, as shown in Table I below.

TABLE I. POTENTIAL SENTENCE PATTERNS LIST

Type	Potential Sentence Patterns	Examples
1	spatial relationship + geographical name + POI's name/category/attribute	北边花园路的地铁站 靠近北航的饭店
2	geographical name + spatial relationship + geographical name	北航南边的北邮
3	geographical name/POI's name/category + POI's attribute	出版社的电话 北航的食堂
4	geographical name + geographical name	北京航空航天大学体育馆 北大的未名湖
5	geographical name /POI's name/category + functional vocabulary + POI's attribute	北航最好吃的餐厅 海淀区便宜的商场
6	geographical name /POI's name + spatial relationship + functional vocabulary + POI's name/category/attribute	北航东边好吃的餐厅 北邮附近好玩的游戏厅
7	geographical name /POI's name/category + spatial relationship	北航的东边 北航的附近
8	spatial relationship + functional vocabulary + POI's name/category/attribute	东边好玩的地方 南边便宜的饭店
9	functional vocabulary + POI's category/attribute	好玩的地方 类似沃尔玛的超市
10	geographical name/spatial relationship/POI's name/category/attribute + functional vocabulary	东边好玩的 北京的好吃的

B. Obtaining Segmentation Sequence and Abstract Information

1) Obtaining segmentation sequence

The question is composed of a continuous string of Chinese characters, and there are no delimiters between words, so segmentation is the foundation of component identification and segmentation merger. For example, the input is "北京航空航天大学附近的肯德基", so the output is('北京', '航空', '航天', '大学', '附近', '的', '肯德基').

There are some Chinese segmentation analyzers, which are LTP of Harbin Institute of Technology, ICTCLAS of Chinese Academy, Paoding, IKAnalyzer, etc. We adopt the LTP system.

2) Obtaining abstract information

In this paper, we need two kinds of abstract information: POS tagging sequence which is HMM's classification object, and dependency relationship sequence which is important basis of component identification and segmentation merger. For example, the input is the question mentioned above, so the output are ('ns', 'n', 'n', 'n', 'nd', 'u', 'nz') and ('ATT', 'ATT', 'ATT', 'ATT', 'ATT', 'VOB', 'ADV', 'DE', 'ATT'), as shown in Fig. 2 below.

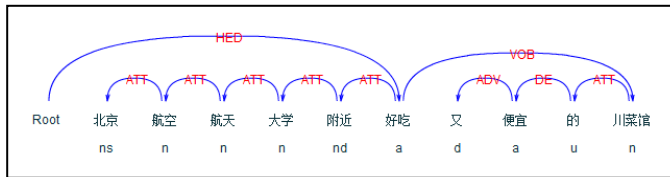


Figure 2. The schematic diagram of dependency relationship and POS tagging

3) Analysis of the results

We can obtain POS tagging sequence and dependency sequence after inputting the segmentation sequence obtained in the previous subsection into LTP. Statistical results show that there are twenty eight kinds of common parts of speech and fifteen kinds of common dependency relationships for natural language query questions about transport travel information, as shown in Table II and Table III below.

TABLE II. POS TAGGING SYMBOLS LIST

Symbol	POS	Symbol	POS
a	Adjective	ni	Organization
b	Other noun-modifier	nl	Location noun
c	Conjunction	ns	Geographical name
d	Adverb	nt	Temporal noun
e	Exclamation	nz	Other proper noun
g	Morpheme	o	Onomatopoeia
h	Prefix	p	Preposition
i	Idiom	q	Quantity
j	Abbreviation	r	Pronoun
k	Suffix	u	Auxiliary
m	Number	v	Verb
n	General noun	wp	Punctuation
nd	Direction noun	ws	Foreign words
nh	Person name	x	Non-lexeme

TABLE III. DEPENDENCY RELATIONSHIP RELATIONSHIPS LIST

Symbol	Dependency Relationship	Symbol	Dependency Relationship
SBV	subject-predicate relation	COO	coordinate relation
VOB	verb-object relation	APP	appositive relations
ATT	attributive + verb structure	VV	serial verb construction
DE	the construction about "的"	IS	absolute construction
HED	core word	IC	independent clause
QUN	quantitative relation	POB	preposition -object relation
ADV	adverbial-verb structure	WP	punctuation marking
LAD	front additional relation		

C. Establishing the Sentence Pattern Classifier

Establishing the sentence pattern classifier is the training process of HMM parameters, whose final results are π , A and B defined in section II. Questions can be trained in three approaches with HMM. The first one is to take original visible symbols as training samples, which is difficult to achieve because too many original symbol types will lead to excessive matrix dimension. The second one is to take dependency relationship symbols as training samples, whose accuracy is relatively poor because relatively fewer dependency relationship types will cause inaccurate prior probability; the third one is take POS tagging symbols as training samples the paper adopts.

Taking the convenience of programming into account, we define the hidden state and the visible symbol as POS, so A and B are defined as 28-order square matrix, according to the statistical results of the previous subsection. Determining the value of transition probability a_{ij} and $b_j(k)$ is learning issue of HMM. First of all, these questions collected are subjectively classified, combination of potential sentence patterns. And then, these POS tagging sequences are input into HMM, and the HMM calculates the value of a_{ij} saved in the matrix A and $b_j(k)$ saved in the matrix B, by the famous baum_welch algorithm [13,14,15]. The specific training steps are as follows:

- Initializing the model $\lambda = (\pi, A, B)$;
- Calculating forward probability $\alpha_t(i)$ and backward probability $\beta_t(i)$;
- Calculating estimators $\hat{\pi}_i$, \hat{a}_{ij} and $\hat{b}_j(k)$ of π_i , a_{ij} and $b_j(k)$ by revaluation formula;
- Calculating the probability $P(O/\hat{\lambda})$, where $P(O/\hat{\lambda})$ is the conditional probability of producing the POS tagging sequence $O = (o_1, o_2, \dots, o_t)$ by the model $\hat{\lambda}$. We return the step b), till the discrepancy of $P(O/\hat{\lambda})$ and $P(O/\lambda)$ is meet the precision.

According to life experience, we get more than 2300 sentences about POI query from the major review websites, traffic and tourism websites, search engines, and open questionnaire. After removing repeated and ineffective sentences, there are 1849 syntactically correct sentences, which can be classified into the first six types of potential sentence patterns obtained by previous preprocessing, based on the appearance order and semantics form of four key components. We take 1649 sentences as training samples. Each HMM matrix trains less than 300 natural language questions, in which the four types of vocabulary have respectively changed. Offline training time of each sentence pattern is about 12s. Finally 6 groups of HMM $\lambda_i = (\pi_i, A_i, B_i)$, ($i = 1, 2, \dots, 6$) are obtained, however, it is inconvenient that we list them here, because their dimension is large.

D. Classifying Questions to be Tested

The process of classify questions is the valuation process of HMM. Question's POS tagging sequence is input into the six groups HMM $\lambda_i = (\pi_i, A_i, B_i)$, and each HMM λ_i calculates an probability value which the sentence pattern

corresponding to λ_i produces the sequence by the forward algorithm [13,14]. The specific training steps are as follows:

- Initializing forward probability:

$$\alpha_t(i) = \pi_i b_i(o_1), 1 \leq i \leq N \quad (4)$$

- Iteratively calculating forward probability:

$$\alpha_{t+1}(j) = \left[\sum_{i=1}^N \alpha_t(i) a_{ij} \right] b_j(o_{t+1}), 1 \leq t \leq T-1, 1 \leq j \leq N \quad (5)$$

- Calculating the conditional probability:

$$P(O/\lambda_i) = \sum_{i=1}^N \alpha_T(i) \quad (6)$$

Comparing the six probability values, the sentence pattern corresponding to the largest probability value is most likely to produce the sequence, so the question corresponding to the sequence is classified to the sentence pattern.

E. Extracting the Four Kinds of Key Components of Questions

In the section, we give the mathematical description of the issue of obtaining key components and a complete example which shows the whole process of test question processing.

1) Problem description

We assume that θ is the space constituted by the natural language vocabulary, then the dependency relationships sequence $R = (r_1, r_2, \dots, r_k)$ is a vector of θ ; the space σ is constituted by the vector $M = (NS, ND, FUNC, TYPE)$. In order to solve component identification problem, essentially, we need find a semantic function F_{cmd} such that $M = F_{cmd}(R)$, whose definition domain is θ and whose value range is σ .

The vector M^* can be output when R is input into the computer system F_{cmd}^* , which is constructed by the existing dependency parser. The correctness of the F_{cmd}^* can be judged by comparing the discrepancy of the output and true value in theory, which is $\emptyset = ||M|| - ||M^*||$. If $\lim_{n \rightarrow \infty} \emptyset = 0$ is confirmed statistically, F_{cmd}^* is effective.

2) A complete example shown

The example is processed based on the premise that the HMM has been trained. The test question is "北京航空航天大学附近好吃又便宜的川菜馆有没有". Firstly, a relatively declarative sentence"北京航空航天大学附近好吃又便宜的川菜馆" is obtained by filtering out tone composition of question; secondly, the LTP system gives segmentation sequence, POS tagging sequence and dependence relationship sequence of the declarative sentence, as shown in Fig. 3 below:

```
<sent id="0" cont="北京航空航天大学附近好吃又便宜的川菜馆">
  <id="0" cont="北京" pos="ns" parent="1" relate="ATT" />
  <id="1" cont="航空" pos="n" parent="2" relate="ATT" />
  <id="2" cont="航天" pos="n" parent="3" relate="ATT" />
  <id="3" cont="大学" pos="n" parent="4" relate="ATT" />
  <id="4" cont="附近" pos="nd" parent="5" relate="ATT" />
  <id="5" cont="好吃" pos="a" parent="1" relate="HED" />
  <id="6" cont="又" pos="d" parent="7" relate="ADV" />
  <id="7" cont="便宜" pos="a" parent="8" relate="DE" />
  <id="8" cont="的" pos="u" parent="9" relate="ATT" />
  <id="9" cont="川菜馆" pos="n" parent="5" relate="VOB" />
</sent>
```

Figure 3. The result of segmentation sequence and abstract information

Among them, these symbol 'id', 'cont', 'pos' and 'parent' denote respectively segmentation's order, Chinese segmentation, POS, parent node, dependency relationship; thirdly, POS tagging sequence is input into the HMM trained in the previous subsection, and the HMM gives the most likely sentence pattern is the Type 6; finally, according to the order of four key components in the sentence pattern and dependency relationship sequence, the four segmentations ('北京', '航空', '航天', '大学'), which the four "ATT" in front of "nd" correspond to, are merged as NS "北京航空航天大学" by the order of parent, and the segmentation '附近' "nd" corresponds to is identified as ND, and the segmentations ('好吃', '又', '便宜') in front of "DE" are identified as FUNC which is divided into two parts by the segmentation '又' "d" corresponds to, and the segmentation '川菜馆' "VOB" corresponds to is identified as TYPE.

IV. EXPERIMENT RESULT

In this section, the key component extraction method the paper proposes is tested. This test includes questions classification test and key component extraction test. Test data is 200 questions of the 1849 questions mentioned above.

We define the accuracy rate ρ and the recall rate δ , as follow:

$$\rho = \frac{num_1}{num_2} \quad (7)$$

$$\delta = \frac{num_1}{num_3} \quad (8)$$

Where num_1 is the number of questions which are correctly classified or identified for some sentence pattern, num_2 is the number of questions which are classified to the same sentence pattern, num_3 is the number of questions which should be correctly classified or identified for some pattern.

A. Questions Classification Test and Analysis

Questions classification is viewed as the essential component of key component extraction, so its accuracy will affect the method's performance. In order to ensure each HMM to have the same geographical name and spatial relationship, we manually replace them in each sentence pattern, which make test data in the same group satisfy the fairness.

For the six most commonly used question types, Fig. 4 illustrates the result. As Fig. 4 shows, each HMM's accuracy rate is more than 90% and the accuracy rate of questions classification is 94.2%, which proves the effectiveness of HMM classification.

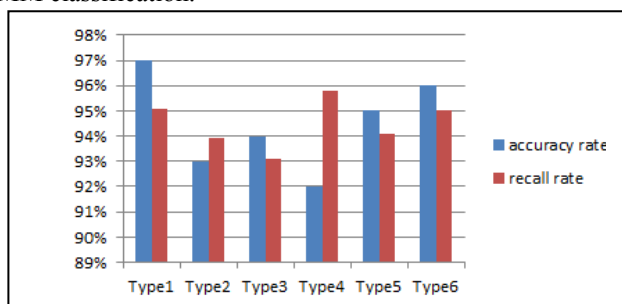


Figure 4. Accuracy rate and recall rate of HMM classification

B. Key Component Extraction Test and Analysis

For these questions classified correctly, we need to further test the performance of key component extraction.

F1-Measure is a commonly used evaluation index in the field of information retrieval and NLP. It is defined as follow [16]:

$$F_1(\rho, \delta) = \frac{2 * \rho * \delta}{\rho + \delta} \quad (9)$$

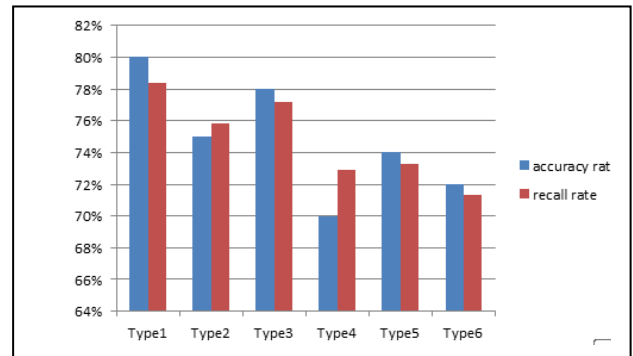


Figure 5. Accuracy rate and recall rate of key componts extraction

As Fig. 5 shows, each HMM's accuracy rate is more than 70% and the accuracy rate of key components extraction is 75.8%. The F1-Measure is 0.83 by (9), which proves the effectiveness of the key component extraction method.

V. CONCLUTIONS AND FURTHER WORKS

The paper presents a key component extraction method for query questions about transport travel information. First of all, six common sentence patterns' characteristic matrices have been calculated from a large amount of real questions. Afterwards, these key components are extracted combination of sentence pattern's structure and dependency relationship. As shown in section IV, the experiments have demonstrated the effectiveness of the key component extraction method, which may be conveniently applied to analyze query questions about transport travel information.

There are two aspects of further works. On the one hand, question's tone composition isn't considered in the method, but part of natural language questions take natural language particle as the main feature, so we need to further process the tone composition to enhance the method's performance. On another hand, on the basis of the paper, the key component should be parameterized, so that the results meeting these conditions are obtained, then they are correctly displayed on the map.

ACKNOWLEDGMENT

The research was supported by Key transportation energy - saving technology based on vehicle networking and applied research (No. 2012AA111903).

REFERENCES

- [1] Yuehua Chen, Laibin Lin, "Application and research on information extraction in natural language query interface", Computer & Digital Engineering, vol. 4, pp. 161-163, 2008. (in Chinese)

- [2] Jinshan Ma, Research on Chinese Dependency Parsing Based on Statistical Methods, PhD thesis, Harbin Institute of Technology, 2007. (in Chinese)
- [3] Wenlin Yao, Research and Implement on Chinese Dependency Parsing, Master's thesis, Ocean University of China, 2009. (in Chinese)
- [4] Jianfeng Gao, H. Suzuki, "Unsupervised learning of dependency structure for language modeling", Proceedings of the 41th ACL, July 2003, pp. 521-528.
- [5] D. Klein, The Unsupervised Learning of Natural Language Structure, PhD thesis, Stanford University, 2005.
- [6] M. Y. Kim, J. H. Lee, "Two-phase s-clause segmentation", IEICE Transactions on Information and Systems, 2005, E88(7):1724-1735.
- [7] Y.Kun, K.Sadao, H.Liu, "A three-step deterministic parser for Chinese dependency parsing", Proceedings of NAACL HLT 2007, Rochester, NY, 2007, pp. 201-204.
- [8] R.Ghani, S.Slattery, Yiming Yang, "Hypertext categorization using hyperlink patterns and meta data", The Eighteen International Conference on Machine Learning, 2001, pp. 178-185.
- [9] Haoyong Xiong, Research and Implement of Chinese Text Categorization Algorithm Based on SVM, Master's thesis, Wuhan University of Technology, 2008. (in Chinese)
- [10] Qingyang Yuan, The Research and Implementation of Chinese Text Classification Technology Based on Decision Tree, Master's thesis, Northeastern University, 2008.
- [11] Zhaohui Guo, Shaohan Liu, Gangshan Wu, "Feature selection for neural network-based Chinese text categorization", Application Research of Computers, 2006, pp. 161-164. (in Chinese)
- [12] Satish L. Gururaj BI, "Use of hidden markov models for partial discharge pattern classification", IEEE Transactions on Dielectrics and Electrical Insulation, vol. 28, pp. 172-182, Apr 1993.
- [13] Shipin Du, Theory of Hidden Markov Models and Its Applications, Master's thesis, Sichuan University, 2004. (in Chinese)
- [14] Richard O. Duda, Peter E. Hart, David G. Stork, Pattern Classification, 2nd ed., New York: Wiley-Interscience, 2000, pp. 105-115.
- [15] Rabiner L, "A tutorial on hidden markov models and selected applications in speech recognition", Proceedings of the IEEE, 1989, pp. 257-286.
- [16] J.D.M. Rennie, Derivation of the F-measure, <http://people.csail.mit.edu/jrennie/writing/fmeasure.pdf>, 2004.

The Role of Teacher in the Multimedia-based Foreign Language Classes

Ketevan Gochitashvili
Language Centre
Ivane Javakhishvili Tbilisi State University
Tbilisi, Georgia

Abstract: The paper focuses on the problem of multimedia-based foreign/second-language classes, functions and advantages of multimedia in the foreign/second language education. The main issue of the present article is to describe the role of teacher in the multimedia-based foreign classes and give them some practical suggestions.

Keywords: Foreign/second language teaching; Language Teacher's Task; Multimedia.

I. INTRODUCTION

Today's achievements in technology influence our every day lives. Information and communication technologies have entered every aspect of our lives. Education, generally and Second/Foreign Language Education, particularly, are not exception as well. That is why number of Electronic and multimedia second language courses have been developed during past decade.

It is certain, that modern students (especially young generation) enjoy learning through computer/technology and teachers and educators have to response their needs and learning requests.

II. MULTIMEDIA AND FOREIGN LANGUAGE LEARNING

Currently using multimedia in educational process is well-known method in teachers' repertoire but the question – Why multimedia based method and what is the role of teacher in the multi-media-based language classes? – Still needs some clear and direct responses.

In the current report we will try to define above mentioned issues and give some practical suggestions as well. They might be helpful for educators due to the fact that they are based on our personal experience besides the existing theoretical researches.

Firstly, we will name advantages of multimedia-based education. From our point of view, the reasons might be as follows:

1. The technology can be major factors for bridging the gap between language learner and target language/culture.
2. Multimedia/computer-based learning process creates cozy and comfortable atmosphere for learners because they are used to use computer/multimedia tools in their everyday

activities and no need for special environment.

3. Interactive computer network allows students to test the result of learning without the risk of being punished for any mistake. Learning does not have to be a pressure. Computer-assisted language learning can reduce the anxiety of students and turns out to be a positive side of learning [1]. VS, they can accomplish their assignments until the answers are correct.

4. It is important that both parties, students and teachers benefit from the outcomes of multimedia-based educational process.

Besides those advantages, we can consider the lack of privacy as a major challenge for multimedia-based method.

Another challenges can be the questions: which multimedia tools work well in language teaching process? How choose appropriate tools?

Technologies which support to language learning are those which allow learners maximum opportunity to interact within meaning-rich contexts through which they construct and acquire competence in the language. Examples of these types of technologies include text-reconstruction software, concordance software, telecommunications, and multimedia simulation software [2].

No doubt, language teaching is a long, tiring and difficult process requiring students' effort, will, time etc. Nowadays student require new tools for "relieving"/simplifying learning process. As they are our main "employers", we should satisfy as many their expectations as possible.

We have experienced, that some teachers and students consider the multimedia as the most important segment in the computer/multimedia based language teaching process. That is a big mistake. There is a simple golden rule we have to keep in our mind: Multimedia is not a massage! Multimedia is just a tool. So, do not be depended just on multimedia tools, it will get you and your student to the failure.

In this paper we will discuss several major practical multimedia tools for the language teaching process which are accessible and easy to use almost for every teacher and learner. It should be stressed, that in spite the fact, that multimedia plays a considerable role in the language teaching, massage (content), as we already mentioned, is most important issue and the multimedia is a just auxiliary means of passing knowledge/content. Activities, both with and without technologies, are fashioned to support active learner use of the target language.

III. MULTIMEDIA TOOLS

E-mail

E-mail, a form of asynchronous computer-mediated communication, has been called "the mother of all Internet applications" [3]. Since the evolution of networks, computers can offer foreign language learners more than drills: "they can be a medium

of real communication in the target language, including composing and exchanging messages with other students in the classroom or around the world" [4]. Indeed, foreign language teachers are just beginning to sense the impact this medium is having on their profession, through the careful examination and creative integration of this tool into their classes.

From my personal teaching experience, using e-mail develops foreign language students writing skills perfectly and it is a really well-known way for teachers. I would describe it as a support my students use the word processor for their written work. Also, besides grammar rules, vocabulary, reading, writing, listening and speaking skills cultural component is a vital part of second/foreign language. Email is a good possibility to teach students language etiquette (register): how to greet, introduce, address, say good-bye to people from different age, social position, cultural groups which is the foundation for effective communication. At the same time, emailing is the part of their real lives as well, so it will be very practical assignment for learners.

E-mail is a perfect way for students' reading and vocabulary improvement as well. Teacher's responsibility is to choose and share proper message and text for his/her students. At the beginning it can be just short/brief information about classes, meetings, schedule (great majority of the words can be expressed by numbers, are familiar or easy to guess for students). Step by step more complicated words and messages can be sent.

What is teacher's job?

1. E-mail can be used from the very beginning of the teaching process. Be careful in choosing message and language repertoire. For the beginners dates, numbers, proper names and some international words (like presentation or training) are suitable.
2. Feedback is must. Any question, response etc. needs adequate feedback (sometimes using student's native language or the language you, teacher and student share (blended with target) can be helpful.
3. Think about corrections. May be immediate corrections discourage students. But leaving them without reaction does not work well. Anyway, correct mistakes and errors personally. Publicly you can just prize the students.
4. Use (but not overuse) symbols. Integrate them in the text (narrative). They are international and can serve as hints for students. It is truly motivating.

Audio and Video Media, Film/TV

Foreign language theorists and teachers suggest foreign/second language tutors to use audio and videotapes in language classrooms for many years now. At the same time, issue can be arguable. Some specialist admit, that they train mainly receptive skills thus risk encouraging a passive attitude in the students, if these media are not used with a clearly defined pedagogical methodology [5]. Yes, that is truth, but receptive skills, like listening, are as important as any others. These audio-video media creates great opportunity to get familiar with native speakers spelling/intonation, gestures, what

is almost impossible without it. Teachers must be careful about good balance. Ignoring the roll of video and audio as a media in the foreign language teaching is a big mistake. TV programs, movies, documentary are good possibilities. It contains both, text and visual materials and each of them supports students' better understanding.

Film and television segments offer students an opportunity to witness behaviors that are not obvious in texts. Film is often one of the more current and comprehensive ways to encapsulate the look, feel, and rhythm of a culture. Film also connects students with language and cultural issues simultaneously such as depicting conversational timing or turn-taking in conversation. For our personal teaching experience, students achieved significant gains in overall cultural knowledge after watching videos from the target culture in the classroom. Culture is not a static phenomena. In movies and TV programs all contemporary customs, behavior, way of expressing feelings (verbally and non-verbally) are illustrated. Elearning and multimedia based language courses have advantage to integrate pieces (or even full versions) of movies and TV programs for their purposes. As we know, language changes are reflected in writings at least after 20-30 years later. As for movies and TV, they react immediately on language changes. It makes clear the importance of movies and TV programs in language learning process [6].

What is teacher's job?

1. Teacher should be very careful in choosing the materials. Some movies, TV programs are appropriate for second language education, some are completely useless.
2. Prepare students; provide them with glossaries, cultural background, historical and socio-cultural knowledge.
3. Of course, you can use students' native language or the language the group members share if it necessary for instructions or explanations (mostly with the beginners, but decrease it as soon as possible). Keep balance - do not overuse familiar languages. It might be comfortable for students but does not provide their progress in new language.
4. To support the reproducing skills through the verbalization (in written or oral form) of the screen materials is the best decision. Ask students to verbalize using the questions below:
 - a) Time and space (when and where the takes place).
 - b) Main characters and supportive characters.
 - c) The meaning and importance of the typed event (on the local, regional, international level).
 - d) Community and society - life style, values, customs, daily life, etc.
 - e) Similarities and differences between learners' native and target cultures.

The number of questions can be extended or reduced due to the necessity.

Social Network: Chat rooms/Chat, Skype, Facebook

As a language teacher, "chatting" is one of the central parts of my classes. It means conversing for meaningful purposes in target language. In face to face classes teacher carefully guides

these “chat sessions” between and among students to enable as much effective, learner-centered interaction as possible. Online (both, synchronous and asynchronous) needs some more gaudiness. It is a “live” process and nobody can be sure about the consequences. Teacher should plan chat sessions in details (where she/he is supposed to interrupt, how to change inappropriate communication, how to encourage students talks etc.).

What is teacher’s job?

1. For this approach to be successful, teacher explains his/her role (sets time, topic of conversation (possible together with students as well);
2. Tutor defines the role of his/her students and the goals of chat.
3. As chat is a part of whole learning process it must be evaluated as any other activities. Teacher develops the evaluation system (she asks for summary of chat; counts how many words or phrases each participant has used and so on, mistakes and correct forms).
4. Besides the communication teacher asks his/her students their conversations systematically reported in written form following a standard written guide. Learners therefore become proficient in reflecting on their experiences and producing written summaries of their chats. These can include 1) where the conversation took place; 2) with whom (how many participants); 3) what was discussed; 4) how long it took place; 5) what is the conclusion; and 6) how the individual felt about the interaction.

The best thing is to engage students with native speakers’ chats. Language teacher encourages his/her students to be concerned less with accuracy and more with what they want to say, what is meaningful for them, and what, as she says, “comes from their hearts” [2].

Video/teleconferences

Videoconferences are a great success and as valuable, as any other written or oral communications for the students on intermediate level. They can improve reading and listening skills and cultural awareness as well. Language teacher should organize videoconferences in target language; any student is welcome to send their contribution in target language. The question of translation can be agreed between conference participants. From my personal experience, in some cases it works good, but sometimes it is not necessary, it depends on conference topic and participants’ language awareness level.

What is teacher’s job?

1. The themes of videoconferences should be announced in advance, so the participants have time to gather relevant material. This can be novels, stories, articles, audio- and videotapes which we study in class and discuss like we do with any other material in the curriculum. Then the students are divided into groups, each with a particular aspect/topic to write about. They use the
2. word processor and thus have time to structure as

well language as contents, before I gather the finished letters into one file open the line and send the whole file.

3. Evaluate students’ job. During the conference, which can last from one single day to several weeks, we can read the contributions from the other participants. This is of course exciting for the students, and evaluations show a noticeable improvement in their reading skills deriving partly from the fact that they read the preparatory material very intensely, because they have to tell others about it, partly from reading a number of letters in the target language from the other participants.
4. Teacher should encourage students’ progress in language fluency. Day by day, assignment by assignment conference topics can become more complicated and wide. Any successes must be prized by teacher and fellow conference participants.
5. Any ideas should be accepted. Just minor corrections (objections) can be done. The role of teacher should be as limited as possible. It is a student’s task, not teachers.
6. Summarize the results of the conference. Any positive aspects should be admitted by the teachers.

At the end of our paper we should touch three basic lines to fulfill multimedia-based foreign language education: first, training and support for teachers of foreign language teachers, provided by specially educated trainers; second, setting up the necessary infrastructure (fully equipped computer labs, connected to the Internet and technical support); third, development of new software and localization or adaptation of existing, international, exploratory and multi-disciplinary educational software.

References

- [1] B. Gates, N. Myhrvold, P. Rinearson, “The Road Ahead.” Penguin Books; Revised edition November 1, 1996, p.32.
- [2] M. Warschauer, C. Meskill, “Technology and Second Language Teaching and Learning”, <http://www.albany.edu/etap/faculty/CarlaMeskill/publication/mark.pdf>.
- [3] M. Warschauer, “Comparing face-to-face and electronic discussion in the second language classroom.” CALICO Journal, 1995, 13(2 & 3), pp. 7-26.
- [4] R. Oxford, Language learning strategies. New York: Newbury House.1990, p.79.
- [5] L. Kornum, Foreign Language Teaching and Learning in a Multimedia Environment, <https://calico.org/memberBrowse.php?action=article&id=549>.
- [6] K. Gochitashvili, “Intercultural Aspects in the Second Language teaching Process in The Bilingual Classes.” Issues of State Language Teaching, Problems and Challenges”, Tbilisi, Batumi, 2012; p.160.

The Educational Communities' social network

Instructor : Khayyam H. Masiyev¹

Students : Nargiz Bayramova², Elvira Siraczade³

Qafqaz University, Baku, Azerbaijan

¹xmesiyev@qu.edu.az, ²nargiz.bf@gmail.com, ³esirajzada@gmail.com

Abstract— Distance education is as a primary instruction which is significantly increasing around schools and universities. Whereas the social networks like Facebook, LinkedIn, MySpace is worldwide used by students in all education centers. The problem is that, the Universities, Schools and other education centers must be interrelated in one system. Hereinafter we will call this system as Educational community's social network. The method which we will use while realizing our idea is to understand current situation, analyze all inputs and outputs, and make cost-benefit analysis and considering on right solution. At the end we will organize a system like will connect all centers with each other it will reduce time-consuming, will increase benefits for anyone.

Keywords— social networking, social media, educational networking, Edunet, Edu-center

INTRODUCTION

Social networking sites continues to grow, educators are seeing their potential for use in education, realizing that social networking sites may have the ability to promote both active learning and collaboration. In fact, Selwyn (2009) claims that social networking may “benefit learners by allowing them to enter new networks of collaborative learning, based around interests and affinities not catered for in their immediate educational environment”. Thus, social networking sites may provide a forum for extending the traditional classroom and enabling users to join groups that match individual educational interests.

1. The current state of social networks

Social networking is one aspect of social media, where individuals are in communities that share ideas, interests, or are looking to meet people with similar ideas and interests. Currently, the two most popular social networking communities are Facebook and MySpace. This guide will focus on the possible uses of Facebook, MySpace, YouTube, Flickr, blogs, Twitter, and del.icio.us for marketing in higher education.

Facebook is a social utility that connects people with friends and others who work, study and live around them. People use Facebook to keep up with friends, upload an unlimited number of photos,

share links and videos, and learn more about the people they meet.

Facebook is made up of six primary components: personal profiles, status updates, networks (geographic regions, schools, and companies), groups, applications and fan pages.

MySpace is an online community that lets you meet your friends' friends, share photos, journals and interests. Because of the way Facebook was started, it has developed an “elite” image, and is more attractive to colleges and universities to adapt, over MySpace.

YouTube is the leader in online video, and the premier destination to watch and share original videos worldwide through the Web. It allows people to easily upload and share video clips across the Internet through Web sites, mobile devices, blogs, and e-mail. Universities have been making videos for 20+ years to aid in recruitment efforts.

Flickr is an online photo site where users upload photos that can be organized in sets and collections. Public photos may be viewed and commented on by others.

Blogs are a form of online journal. They can have a single author, or several. Most blogs allow readers to post comments in response to an article or post.

Blogs are also being used by some colleges to post news articles to open conversations about them. Faculty to blog about their teaching, travel and research. Admissions counselors' blog about their travel and recruitment cycle. [2, p. 4]

Twitter is a cross between instant messaging and blogging that allows users to send short (140-character) updates. Users can also follow the updates of friends they “follow,” send them direct messages, reply publicly to friends, or just post questions or comments as their current status.

del.icio.us, now also available at delicious.com, is one of many social bookmarking Web sites. The primary use of del.icio.us is to store bookmarks online, which allows users to access the same bookmarks from any computer and add bookmarks from anywhere, too. Tags are used to organize and remember bookmarks, as compared to folders built in to Web browsers bookmark tool.

2. The current state of educational network

"Educational Networking" is the use of social networking technologies for educational purposes. The Education Network is learning and teaching resource providing schools with a secure network designed and maintained by experts within the educational community. A dedicated education network, it harnesses the power of broadband technology in order to provide unique content and services, delivering a personalized learning experience in the classroom and enabling users to share learning resources at every level. Optimized for data-intensive applications (including Video Conferencing), The Education Network provides a number of unequalled advantages for schools, offering a secure and safe environment where issues such as copyright are managed and where teachers, pupils and parents can work confidently together.

The Education Network provides a wealth of services and resources, many of which can only be accessed via The Education Network connection. The Education Network works with industry and the government to implement and raise standards for the benefit of learners. The Education Network is the impartial advisor for schools and can also achieve huge savings from aggregated procurement for the benefit of school budgets.

Social media comprises of activities that involve socializing and networking online through words, pictures and videos. Social media is redefining how we relate to each other as humans and how we as humans relate to the organizations that serve us. It is about dialog – two way discussions bringing people together to discover and share information.

I. USES FOR EDUCATIONAL NETWORKS

- Create an environment that cannot be duplicated elsewhere for networking students that will not meet face to face.
- Teachers retain administrative control (ban users, approve photos and videos, make the site public or private, add gadgets).
- Students quickly learn to use such sites which give them a way to blog, share photos, share videos, join groups, and comment and rate the work of one another in peer review.
- Private educational networks can be used to educate students on social networking safety in a classroom setting before they move personally to major platforms like Facebook and Myspace
- Organizations can create communities of learners that last beyond the span of a course, a grade, or even beyond graduation.

II. OBSTACLES TO EDUCATIONAL NETWORKING

- Most people use the word "social" networking which denotes "play" and not classroom professionalism to many educators.
- Use of "embedded" social networking platforms - some are using Facebook and Myspace as their classroom social networks causing an overlap between a space for friends and a space for classrooms.
- Some technology specialists state that teachers have not asked for the technology in their classrooms.
- Perception that social networks are not used by teachers, only by youth.

III. CRITICISMS OF EDUCATIONAL NETWORKING

- The echo chamber effect of only networking with those of similar opinions/attitudes.

3. Edunet learning platform and its key features

Edunet is a learning platform aiming to improve training and education.

a) Edunet provides:

1. a high-quality learning experience for children and young people
2. opportunities for extended and enriched learning
3. encouragement for independent and collaborative learning
4. improved management of teaching and learning
5. possibilities for improving pupil-students attainment and school-university standards
6. improved home-school links and Moodle educational portal
7. single sign in to a range of online services
8. a wealth of on-line tools for creating and using learning resources
9. a key element to our 21st century school program and university degree

Edunet integrates with Information Management System (IMS) allowing teachers, pupils to remotely access key school data including: 1) Pupil, student and reports 2) Assessment data 3) Attendance information

Pupils, students and teachers can access their Council-provided data storage at home using Edunet.

Furthermore, Dropbox and Google Docs integration allows for easy import/export of files between all of these services.

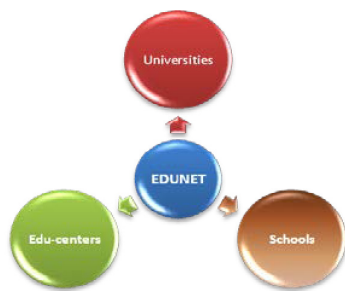
EduNet provides seamless integration with a wide range of appropriate cloud services including Google Docs, Google Apps for Education, Dropbox and Twitter.

4. The structure of EduNet

EduNet is a virtual community for those involved in education in Worldwide. We want to offer all schools and universities Worldwide the opportunity to be listed and the ability to join in the benefits of this Educational Resource.

EduNet consists of 3 parts:

1. EduNet for universities
2. EduNet for schools
3. EduNet for edu-centers

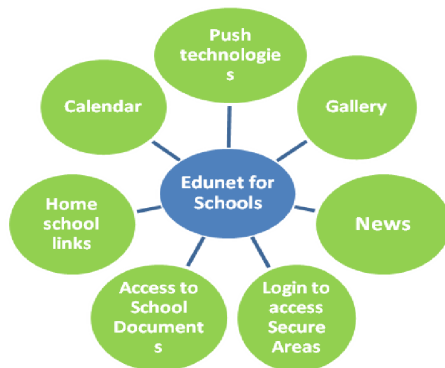


Picture 1. The structure of EduNet

1) Schools can create sites featuring online calendars, picture galleries and a wide range of school documents including school prospectuses, school policies and important school forms.

Latest news articles can be added directly to schools' Web sites or even integrated with schools' Twitter accounts.

All helping to provide an engaging Web presence whilst also pushing communication to parents, and enhancing home school links.



Picture 2. School's EduNet system

2) Universities Information Management System (UIMS) data

Attendance, assessment and behavior data is all available together with the ability to complete reports online.

Moodle

A course management system which allows teachers to create media-rich, engaging courses for their lessons.

Room and Resource Booking

Rooms and resources can easily be reserved by teachers.

My Drive

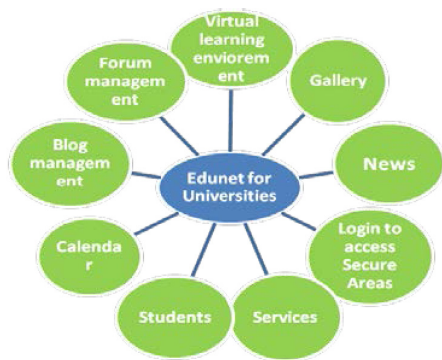
Easy access to documents, not only files and folders stored in school but documents stored using cloud services such as Dropbox and Google Docs.

Administrative Blog Management allows you to:

- Create multiple community and group blogs
- Allow for members to maintain personal blogs
- Browse blogs and posts by tags
- Search blogs and posts by post content
- View, edit and delete blog posts and comments
- Lock specific blogs to member-only access
- Lock specific blogs to read-only access
- Set admin approval on group and community blogs

Forum Management allows you to:

- Create multiple community and group forums
- Search forums and thread content
- View, edit and delete threads
- Lock specific forums to member-only access
- Lock specific forums to read-only access
- Assign specific members as forum moderators
- Create automated forum digests that email members updates



Picture 3. University's Edunet system

3) Edu-centers can be:

- Universities or related research institutes
- Courses
- Training centers
- Industry (automotive, electronics, information technology, telecommunications, civil construction, etc)
- Users (automobile associations, etc)
- Industry federations
- Consultants
- Others (e.g. private research institutes)

Summarizing all educational networks, you will see their key features below:

Key features	Facebook	Twitter	Edunet	Youtube	Skype
Photo sharing	x	x	x		
Video sharing	x	x	x	x	
Online books			x		
Online communication	x	x	x		x
Online lessons			x		
Status update	x	x	x		
Chat conversations	x	x	x		x
Group meetings			x		
Special Interested Groups	x	x	x		
News	x	x	x		
Gallery	x	x	x		
Documents			x		
Online assessment			x		
Web meetings			x		

Table 1. Educational social network's key features

Conclusion

Educational networking is allowing educators to both learn things — which traditional professional development has always afforded them.

- ✓ Educational Network is a social network that promotes Teaching, Learning and Research in Higher Education Institutions, in schools.
- ✓ Educational Network aims to interconnect all the Schools, Universities, Tertiary and Research Institutions and edu-centers by setting up a cost effective and sustainable private network with high speed access to the global Internet.
- ✓ Educational Network also facilitates electronic communication among students and faculties in member institutions, share learning and teaching resources by collaboration in Research and Development of Educational content.

And Educational Networking Takes a Major Step Forward:

1. **First**, collaborating in real-time.
2. **Second**, the ability to easily search out and connect with those who share common needs and interests.
3. **Third**, today's educational networking sites have a hit or miss approach to organizing and sharing content.

REFERENCES

[1] Fort Lauderdale, "Social Networking Pioneer Steve Hargadon Helps Build Unique Online Community for Educators", Florida — June 23, 2009 (p. 1-6)

[2] Rachel Reuben, "The Use of Social Media in Higher Education for Marketing and Communications: A Guide for Professionals in Higher Education", 10 August 2008

[3] Kevin P. Brady, Lori B. Holcomb, and Bethany V. Smith North, "The Use of Alternative Social Networking Sites in Higher Educational Settings: A Case Study of the E-Learning Benefits of Ning in Education by Carolina State University", Journal of Interactive Online Learning Number 2, Summer 2010

[4] Pollara, P & Zhu, J., "Social Networking and Education: Using Facebook as an edu social Space. In Proceedings of Society for Information Technology & Teacher Education", International Conference 2011 (p. 3330-3338)

[5] Federica Oradini, Gunter Saunders "The use of social networking by students and staff in higher education (Online Learning Development)"

[6] Adam Goldfarb, Natalie Pregibon, Jonathan Shrem, Emily Zyko, "Informational brief on social networking in education", February 2011 (p 2-6)

Cognitive methodology to develop a recovery strategy for the sturgeon stocks of the Caspian Sea

R.A. Karayev¹, A.I.L. Payne², N.Y. Sadikhova³, K.A. Aliyev⁴, A.N. Gasimli⁵

^{1,3}Institute of Cybernetics of the National Academy of Sciences of Azerbaijan, Baku, Azerbaijan

²The Centre for Environment, Fisheries and Aquaculture Science (CEFAS), United Kingdom,

⁴Oceaneering International Services LTD, United Kingdom,

⁵Azerbaijan State Oil Academy, Baku, Azerbaijan

karayevr@rambler.ru¹, andy.payne@cefasc.co.uk², natella5@rambler.ru³, kaliyev@oceaneering.com⁴, arazbakuvi@yahoo.com⁵

Abstract - The situation of sturgeons in the Caspian Sea, which many experts currently assess as "catastrophic", is discussed. The untenable ness of traditional approaches to the issue of recovery of stocks in the complex conditions of region developed after disintegration of Soviet Union is substantiated. A new approach to this issue, based on the paradigm of "strategic planning" and "cognitive modeling", is considered as an alternative. A cognitive methodology to develop a recovery strategy for sturgeon stocks is proposed.

Keywords: *Caspian sturgeons, recovery strategy, cognitive methodology*

INTRODUCTION

The massive decline in sturgeon catches in the Caspian Sea, a consequence of anthropogenic and natural causes, needs to be turned around by implementing a management system that has as its primary aim the recovery of the stocks to sustainable levels. The FAO (2004) stresses that "creation of a document giving a strategy for stock recovery should be a priority task for everyone related to the management process of the sturgeon fisheries of the Caspian Sea".

In report the issue condition is analyzed. Statement of a task to develop of recovery strategy, reflecting a modern paradigm of strategic planning is offered. The methodology of the decision of the task, based on ideas and methods of cognitive modeling actively developing in the modern theory of management by ill structured systems is offered

THE ISSUE CONDITION

Recent history has shown that to do this with the help of the traditional tools of fisheries management, such as total allowable catches (TACs), including its modern equivalent, the "precautionary approach", quotas, and technical measures, is not effective. Moreover, to rely on such tools under certain conditions is high risk for the resources.

The solution in the authors' opinion is to invoke management tools at a higher, strategic level. These need to take into account the long-term prospects and targets for the stocks and all system issues related to the problem currently: the status and distribution of the different stocks, the extent of poaching, pollution of the sea and the rivers of the Caspian basin, national and international markets for sturgeon product, national and international legislation, the efficiency of conservation and control measures for the stocks, the developing technologies of artificial production and culture, fluctuations in the sea level of the Caspian, natural climate change, possible geo-conflicts and international terrorism, competitors for food (*Mnemiopsis leidyi*), etc.

Such tools do not exist at present in most fisheries management regimes. Numerous attempts internationally to use strategic planning for fisheries, as applied in the theory of corporate management [1], have tended to fail because they do not take into account:

- the ecological and biological nature of fish stocks that distinguishes them from industrial and economic entities;
- the great uncertainty and risks inherent in fisheries management, which generates unjustified illusions and strategic errors;
- the multicriteria nature of fisheries, which demands that any strategy be based at least on issues related to biological, economic, and social efficiency.

Today it is clear that, despite the appeal of applying a strategic planning (SP) methodology to the management of fish stocks, it would need to be modified on the basis of an integrated approach that synthesizes population biology and ecology with the theories of corporate planning and complex systems, so stressing the principles of and methodology applied to ecosystem processes, experimental verification of which is virtually impossible. Here we

attempt to create a methodology for SP, based on such a synthetic, integrated approach. We offer a general formulation of a SP task, describe the substantive provisions of a SP methodology, and discuss opportunities for its practical application in the Caspian Sea.

SP TASK FORMULATION

In a general view SP task formulation can be presented as follows:

$$\begin{aligned} S: S^* &\rightarrow C(T; B; F); \\ K_S^* &(E_b, E_e, E_s) \rightarrow \text{extr}(max)|_{\Omega, L, N}; \\ R_S^* &\leq R_0, \end{aligned}$$

where S is a set of allowable strategies for recovering a stock; S^* is the "optimal" strategy; $C(T, B, F)$ is the concept of stock recovery (enhanced production); T is the horizon of planning; B represents the biological characteristics of a stock that determine the parameters of reproduction, growth, and mortality; F refers to environmental influences; $K_S^*(E_b, E_e, E_s)$ is the generalized efficiency parameter for strategy S^* , including biological (E_b), economic (E_e), and social (E_s), criteria; Ω is a set of factors of uncertainty caused by incompleteness, discrepancy, and unreliability of problem knowledge; L is a vector of resource restriction (time, financial, technological, legislation, etc.); N is a network of safety, establishing borders of environmental parameter change within which the normal conditions for the stock to survive (temperature, salinity, oxygen, toxicity, food production, river flow, area and status of spawning grounds, etc.) are possible; R_S^* is the risk associated with an optimal strategy for recovery; and R_0 is the allowable level of risk.

The SP task concerns a class of complex ill structured problems. Solution of such tasks requires the help of special cognitive methodology that apply theoretical, empirical, and heuristic knowledge of the problem area.

METHODOLOGY OF SP

The structure of the methodology of the collaborative venture is shown in Figure 1. Set against the mechanism of *long-term planning* in fishery theory (the fundamental forecast), based on extrapolation of the past and present into the future, the SP ideology builds a vector of management in a reverse direction. In other words, it builds the present from what is required in the future. Thus, models of the future environment are constructed, the target position of the stocks in these models is decided, and an effective management strategy (scenario) to achieve the target is established.

The methodology includes the following components.

1. The SP concept (its mission and hypotheses), as described by Legeza (1961), Belyayeva et al. (1998), Vlasenko et al. (1999), Katunin et al. (1999), Majnik

and Schwarzkopf (1999), Zagranichniy et al. (2003), Karyuk et al. (2003), Kuznetsov (2004), Baymukanov (2004), Pourkazemi (2005), Payne (2005), Berkeliev (2005), and Karayev (2000, 2003, 2005, 2006).

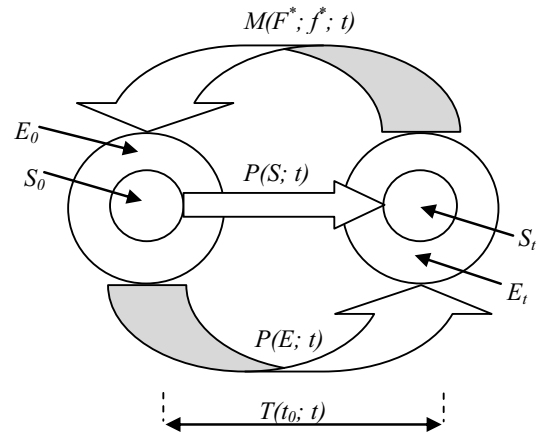


Fig. 1. Logical structure of SP methodology

S_0 – current state of stock; E_0 – current state of the external environment; $T = (t_0; t)$ – the horizon of strategic planning; E_t – future state of the external environment; S_t – future state of the stock; $P(E, t)$ – the procedure for forecasting the state of the external environment; $P(S, t)$ – the procedure for forecasting the state of the stock; $M(F^*, f^*, t)$ – a complex strategy to recover the stock; F^* – managed factors of the external environment; f^* – managed factors of the stock.

2. Conceptual scheme of life cycle of Caspian sturgeon generation (fig. 2)
3. A SP lexicon.
4. An organogram of the SP process.
5. A Reference Model of the strategies (including a Control List of strategic directions and an associated list of the sub-strategies systematized according to the conceptual life cycle of sturgeon) (Appendix).
6. A Control List of efficiency criteria (biological, economic, social) and strategic risks, including rating scales.
7. A Cognitive Map for analysis of problem of recovery of Caspian sturgeon stocks (fig. 3).
8. A Library of the basic procedures of the SP (table 1).
9. A users' manual giving the format of the strategic plan.

MATERIALS

The following materials were used in development of the methodology:

- Declaration on Global Sturgeon Conservation .World Sturgeon Conservation Society. RAMSAR (2006);
- EU/TACIS Sustainable Management of Caspian Fisheries Project (SMCFP) (2004–2006);
- Statistics of legal sturgeon catches in the Caspian basin (1960–2010);
- Data from annual (summer and winter) record surveys in the Caspian Sea (1998, 2000, 2001, 2002, 2004, 2005, 2007, 2009, 2011);

- Collected documents and meeting minutes of the Caspian Sea Commission on Aquatic Bioresources (2011);

- FAO review prepared for CITES (2004);

- Technical Reports issued by the Caspian Environment Program

http://www.caspianenvironment.org/report_technical.htm;

- Draft strategic plans from foreign fishery organizations (web search “fishery strategic planning”);

- Expert knowledge of specialists and scientists of the region, journalistic studies, and observations of fishermen

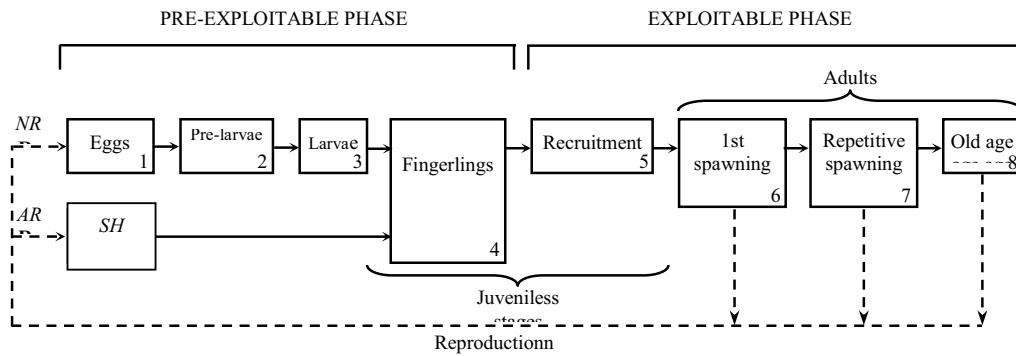


Fig. 2. A conceptual scheme of the life cycle of a sturgeon generation. NR-natural reproduction, AR-artificial reproduction, SH-sturgeon hatcheries

TABLE 1
PROCEDURES AND METHODS OF METHODOLOGY

Procedures	Used methods
P1. Statements of the mission and the strategic purpose, including the target direction	Methods of SWOT- and PEST- analysis of the stocks and their environment [1, 13]. Technical Reports of Caspian Environmental Program (http://caspianenvironment.org)
P2. Search forecasting the long-term behavior of the external environment in the presence of uncertainty	Method of deterministic analysis of the stock-environment system [2]. Foresight-technologies [3]
P3. Construction scenarios of development of environment	Methods of scenario planning [4, 5] Methods of strategic Foresight [3]
P4. Construction Cognitive Map for analysis of problem of recovery of Caspian sturgeon stocks (versions for Volga population of the Russian sturgeon and Kura population of the Persian sturgeon)	Method of construction of Fuzzy Rule-Based Cognitive Maps [5-8, 12] Method of non metric multy dimension scaling [6]. Methods of expert estimation [11]
P4. Generation of alternative strategies (national and regional) for recovering the stock, taking account of scenarios of development of stock and environment	Method of dynamic scenario analysis of Cognitive Maps (Mode of self-development, mode of controlled development) [7, 8]
P5. An analysis of the risks (including biological or economic risks of extinction of populations and an evaluation of risk management strategies)	Method of estimation of risk of the strategy under uncertainty [9, 10]
P6. Multicriteria analysis of strategies and decision-making in conditions of uncertainty and risk;	Method of hierarchy analysis [Saaty T.]
P7. Strategic decision-making support	Method an interpretation of strategic decisions at the level of operational planning [2, 13]

OPPORTUNITIES OF THE METHODOLOGY

The methodology is executed within the format of a

system shell [11], which can be adjusted to address separate populations, or separate areas of the sea (Volga-Caspian, Ural-Caspian, Kura-Caspian, Iranian waters, Turkmenistan waters).

It can also be used to decide on the relevance of

tasks: early diagnostics and forecasting the critical status [10] of the different populations (Russian, Ural, Persian sturgeon); a strategic environmental impact assessment of offshore projects; identification of critical scenarios of economic activities in different areas of the Sea; completion of a Strategic Action Plan for the Caspian Sea (CEP, 2003); strategic diagnostics of the ongoing and the planned projects (e.g. moratorium on fisheries, embargo on trade in caviar and sturgeon products, construction of new hatcheries). In developing the methodology, the results of long-term research of the region's research institutions (CaspNIRKh, Russia; SPCF, Kazakhstan; AzerNIRKh, Azerbaijan; International Research Institute on Sturgeon, Iran; reports of the Caspian Environmental Programme, CEP, <http://www.caspianenvironment.org>; SP projects of foreign fishery organizations through internet searches for "fishery strategic planning"; expert knowledge of the specialists and regional and international scientists; knowledge of fishermen; media, i.e. journalist, investigations) can be used. Correctly and in time, strategic decisions if effected will play a key role in recovering the Caspian basin sturgeon stocks. If successful there will be a future for sturgeon in the Caspian; if they fail, the die is cast.

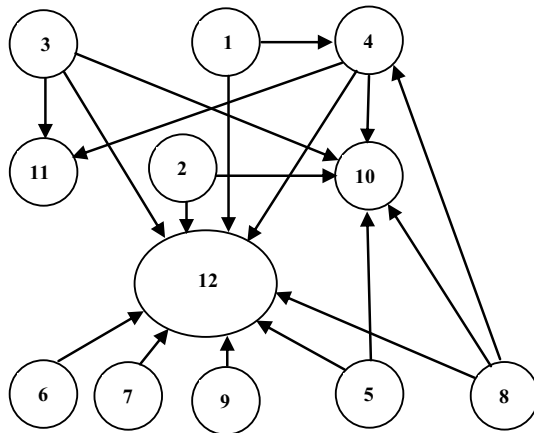


Fig. 3. Prototype of Cognitive map for the analysis of a problem of recovery of sturgeon stocks (The prototype is adjusted on conditions of concrete population of the sturgeon)

Factors of Cognitive map: 1–Sea level, 2–Annual river drain, 3–Hydrological and hydrochemical mode of the rivers and the sea, 4–Forage reserve, 5–Fishing, 6–Deregulation a drain of the rivers, 7–Poaching, 8–Pollution of the rivers and the sea, 9–Disintegration of uniform system of protection and reproduction, 10–Natural reproduction, 11–Artificial reproduction, 12–Stock biomass

APPENDIX

REFERENCE LIST OF STRATEGIES

1. Reproduction of the stock

1.1. Natural reproduction

1.1.1. Reclamation of spawning grounds (dredging operations, reconstruction of fish ladder channels and removal of aquatic vegetation from the river delta)

1.1.2. Creation of artificial spawning grounds

1.2. Artificial reproduction

1.2.1. Construction of an Intergovernmental Reproduction Complex

1.2.2. Technical upgrading of biotechnologies at existing sturgeon hatcheries

1.2.3. Selection of an optimum mass of released fry for a given region of the sea

1.2.4. A sustainable method of releasing fry into the natural environment

1.2.5. Formation of spawning schools and receipt of reproductive products *inter vivo*

2. Protection of stock

2.1. National fish conservation inspectorates

2.2. The use of (combined arms) force against poachers during spawning runs ("Putina" in Russian waters; "Vekre" in Kazakh waters)

2.3. Regional fish conservation inspectorate

2.4. Regional Agreement on the Protection of Fish Resources

2.5. International Conventions on trade in fish products and responsible fishing

3. Protection of the habitat

3.1. Monitoring and documenting spawning grounds

3.2. Monitoring quarter quality of spawning streams

3.3. Monitoring the marine environment (water quality, bottom sediment, soil)

3.4. Monitoring the mutagen city of the marine environment

4. Stock management

4.1. Planning level

4.1.1. Strategic (Strategic planning for a period of up to 30 years)

4.1.2. Tactical (Regional Agreements for 2–3 years)

4.1.3. Short-term (stock appraisal, establishment of TPC, national quotas, and Fishery Regulations)

4.2. Organizational management structure

4.2.1. Private companies

4.2.1. National (governmental) monopolies (along the lines of the "Shilat" organization in Iran)

4.2.2. Regional coordinating agency (Commission on Aquatic Bioresources of the Caspian Sea)

5. Regulatory and legal issues

5.1. National legislation on fishing

5.2. Regional Agreements

5.3. International Conventions (CITES, Bonn Convention)

5.4. European Union law (Berne Convention, the "Natura" list, Attachment B of Council of Europe Directive No. 338/97)

6. Investment and financial strategies

6.1. Bank credits.

6.2. Public funds.

6.3. Private investments by regional and foreign investors.

7. Social strategies (strategies of socioeconomic support for the populations in coastal regions)

8. Risk management strategies

8.1. Development of aqua cultural sturgeon farms in the region

8.2. Import of aqua cultural products from foreign producers (China).

8.3. Creation of a gene pool and cryogenic storage of sturgeon reproductive products

9. Monitoring of the stock.

9.1. Annual surveys takes

9.2. Monitoring of the overall functional state of the reproductive system and of morphogenic attributes of individuals

9.3. Monitoring the physiological and genetic state of the population

9.5. Monitoring the genetic safety of marine products

9.6. Monitoring the control indices of implemented measures

10. The control over a mnemiopsis population

10.1. Introduction of an obligate predator, for example, *Beroe ovata*

10.2. Amplification {strengthening} of a competition with mnemiopsis at a trophy level with the help of active consumers Caspian zooplankton, for example, *Anchousovi kilka*

Modelling, ICCM 2010, 24-26 July 2010.

Philadelphia PA, USA. 564 p. URL:

<http://iccm2010.cs.drexel.edu/call.html>

- [13] Karayev R.A. Sustainable Management of Caspian Sturgeons: Strategic Approach / Proceeding of the UNESCO International Workshop. "Large-scale Disturbances and Recovery in Aquatic Ecosystems". Varna, Bulgaria. 2005.

REFERENCES

- [1] Andrews, K. The Concept of Corporate Strategy. Illinois: Irwin Homewood,. 1987. 453 p.
- [2] Karayev, R.A. Descriptive Models of Sturgeon Population Dynamics / Proceedings of the 5th International Symposium on Sturgeon. Ramsar, Iran. May 2005, pp. 273–275.
- [3] Andy Hines, Peter Bishop. Thinking about the future: Guidelines for Strategic Foresight. N.Y.: Free Press. 2007. 357 p.
- [4] Karayev, R.A., Ismaylov, S.F., and Sadikhova, N.U. Models of the scenario analysis // Transactions of the National Academy of Sciences . 23(3): 33–36, 2003. (in Russian).
- [5] Karayev, R. A. Knowledge-based Models of Ecological Dynamics // Transaction of Academy of Sciences of Azerbaijan, № 2-3: 63-68, 2000. (in Russian).
- [6] Karayev, R.A. 2006. Modelling Caspian sturgeon population dynamics: a new paradigm and new technology // ICES Journal of Marine Science, 63(6): 980–994, 2006.
- [7] Karayev, R.A. Cognitive modelling for decision making in the strategic planning / Proceedings of the IY-th Russian Symposium "Strategic Planning and Evolution of Enterprises". Moscow, April 12–15 2003. Moscow, Central Economics and Mathematics Institute of Russian Academy of Sciences.
- [8] Karayev, R.A., Fuzzy cognitive maps for generation and analysis of managerial decisions. The project awarded with the Diploma of the Russian Academy of Sciences (Branch of Economy) and the Grant of the International Scientific Foundation of Economic Researches of academician N.P.Fedorenko. 2003.
- [9] Dulvy N. K., Ellis, J. R., Goodwin, N. B., Grant, A., Reynolds, J. D. and Jennings, S. Methods of assessing extinction risk in marine fishes // Fish and Fisheries, 5: 255-276, 2004.
- [10] Soule, M. (Ed.). Viability Populations for Conservation. Cambridge University Press. 1987.
- [11] Waterman, D. A Guide to Expert Systems. Addison-Wesley, New York. 1986.
- [12] Proceedings of the 10 th Intern. Conf. on Cognitive

Support Vector Domain Description for non-stationary data

Foued Theljani¹, Kaouther Laabidi¹, Salah Zidi², Moufida Ksouri¹

¹ LACS, ENIT, BP 37, le Belvdre 1002 Tunis, Tunisia

² LAGIS, USTL, Villeneuve d'Ascq, 59650, Lille, France

Email : {foued_theljani, labidi_kaouther, salah_zidi, moufida_ksouri}@yahoo.fr

Abstract—The support vector domain description (SVDD) is an efficient kernel method commonly used in the one-class classification. However, the training algorithm solves a constrained convex Quadratic Programming (QP). This assumes dense sampling in advance (offline training) and requires large memory and enormous amounts of training time. These drawbacks are partially resolved using an incremental SVDD with novelty detection. Despite the improvements above, most of studies don't deal with the paradigm of non-stationary data which exists in many real life applications. In this paper, we propose an incremental SVDD for stationary as well as non-stationary data. The principle is based on the dynamic removal/insertion of information according to well adequate rules. The developed approach is assessed afterwards on some synthetic data to prove its effectiveness.

Index Terms—Incremental SVDD, Non-stationary data, One-class classification, Novelty detection.

I. INTRODUCTION

THE support vector domain description, proposed by [1], [2], is an efficient technique employed to solve one-class classification problems (known also novelty detection problems). The fundamental goal of one-class learning is to generate a rule that distinguishes between a set of target objects called the target class and unseen-novel objects designated as outlier class. To fall into the optimum, the SVDD training algorithm assumes dense sampling and requires all training examples to be available at once for a single "batch" learning step: if a new sample is presented, the classifier must be retrained from scratch. The concern arises when data cardinality increases insofar as the process becomes embarrassing in terms of memory and training time. In the literature, there some works that addressed this dilemma and they resolved in some measure the drawbacks above [3]–[5]. These propositions, based on the Karush-Kuhn-Tucker conditions (KKT) [6], apply the optimization process by taking into account one sample or a limited frame of samples in each run. Furthermore, some approximations are usually adopted on the cardinality of the working data-set in order to ensure the rapidity of convergence during the training process [3], [7]. Accordingly, the data domain is maintained incrementally with respect to samples newly added. Although these improvements are significant, there are still difficulties related to the effectiveness of the SVDD technique when dealing with evolving data in non-stationary environment. Indeed, most of proposed methods optimize the standard deviation of the

Gaussian kernel using a predefined target error rate on the whole data. In this sense, these methods don't support non-stationary data since parameters of non-stationary dataset, such as mean and variance, are not static, but rather they evolve over the time. This may be not adequate for various real life applications such as fault conditions, system monitoring, face recognition, modeling in video, etc.

In this paper, we present a new domain description dedicated for non-stationary data environment. In this framework, we address two main aspects. The first one concerns particularly the issue of data availability. Throughout the optimization process, we consider that training samples are not all available. They may be presented sequentially one by one, which is more proper and realistic. The second aspect investigates the non-stationarity dilemma. We propose a new method based on the dynamic removal/insertion of data according to their contribution in the domain description. Specific rules are employed here to ensure the compactness of data taxonomy.

The paper is organized into four sections. In section.2, we introduce a global overview of the support vector domain description and its theoretical foundation. As for section.3, it is reserved to describe the proposed procedure dedicated to deal with non-stationary data. The results of the experimental tests are exposed in section.4. In the conclusion, we summarize the presented work and we discuss the interest that yields.

II. SUPPORT VECTOR DOMAIN DESCRIPTION SVDD

Let $\chi = \{x_1, \dots, x_i, \dots, x_N\}$ be a target training set, with $\chi \subseteq R^d$. SVDD is a well known one-class or data description kernel machine, which finds a smallest hyper-sphere to contain most of training instances of a target class with some relaxation defined by slack variables. Its original form is formulated as a constrained optimization problem as follows:

$$\min R^2 + C \sum_{i=1}^N \xi_i \quad (1)$$

$$\begin{aligned} s.t. \quad & (x_i - a)^T (x_i - a) \leq R^2 + \xi_i, \\ & \xi_i \geq 0 \quad \forall i = 1, \dots, N, \end{aligned} \quad (2)$$

where R and a denote the radius and center of the hyper-sphere and C is the regularization parameter which gives the trade-off between the volume of the sphere and the misclassification

errors. The variable ξ_i is a slack variable designating the distance of i^{th} data point from the sphere boundary. This is a quadratic optimization problem and can be solved efficiently by introducing Lagrange multipliers for constraints. The Lagrangian formulation of the problem is given thus by the following formula:

$$L = R^2 + C \sum_i \xi_i - \sum_i \alpha_i \left(R^2 + \xi_i - (x_i - a)^T (x_i - a) \right) - \sum_i \gamma_i \xi_i \quad (3)$$

where α_i and γ_i are Lagrange multipliers, $\alpha_i \geq 0$, $\lambda_i \geq 0$. Note that for each training data point x_i , a corresponding α_i and γ_i are defined. L has to be minimized with respect to R , a , and ξ_i and maximized with respect to α_i and γ_i . Taking the derivatives of (3) by setting $\partial L / \partial R = 0$, $\partial L / \partial a = 0$ and $\partial L / \partial \xi_i = 0$, we obtain the Karush-Kuhn-Tucker (KKT) conditions given by the following relations:

$$\begin{aligned} \frac{\partial L}{\partial R} = 0 &\Rightarrow \sum_{i=1}^N \alpha_i = 1, \\ \frac{\partial L}{\partial a} = 0 &\Rightarrow \sum_{i=1}^N \alpha_i x_i = a, \\ \frac{\partial L}{\partial \xi_i} = 0 &\Rightarrow 0 < \alpha_i < C. \end{aligned} \quad (4)$$

The QP equations are obtained by substituting the above KKT conditions in (3). We obtain a dual problem expressed by:

$$\max \frac{1}{2} \sum_{i=1}^N \alpha_i (x_i^T x_j) - \frac{1}{2} \sum_{i=1}^N \sum_{j=1}^N \alpha_i \alpha_j (x_i^T x_j), \quad (5)$$

$$s.t \sum_{i=1}^N \alpha_i = 1, \quad 0 \leq \alpha_i \leq C \quad \forall i = 1, \dots, N. \quad (6)$$

After solving such a standard Quadratic Programming (QP), we obtain the solution $\alpha_i = \alpha^*$, whose corresponding training instances can be classified as Boundary Support Vectors (*BSVs*) outside the hyper-sphere, if $\alpha_i = C$, Non-Support Vectors (*NSVs*) inside the hyper-sphere, if $\alpha_i = 0$ and Non-Boundary Support Vectors (*NBSVs*) just at the hyper-sphere, if $0 < \alpha_i < C$. Any data satisfying one of the above KKT conditions is designated as target class. Otherwise, it is an outlier. Furthermore, all points with $\alpha_i > 0$ are called Support Vectors (SVs) which restrict the data domain and can fully describe the one-class boundary. We can write thus $SVs = BSVs \cup NBSVs$. According to (4), the center a can be easily calculated as:

$$a = \sum_{i=1}^N \alpha_i x_i. \quad (7)$$

To make prediction on an unknown instance z , the squared distance to the center of the sphere must be calculated using the following formula:

$$d^2(z, a) = (z^T z) - 2 \sum_{i=1}^N \alpha_i (z^T x_i) + \sum_{i=1}^N \sum_{j=1}^N \alpha_i \alpha_j (x_i^T x_j). \quad (8)$$

Afterward, we define the decision function as:

$$F(z) = R^2 - d^2(z, a) \quad (9)$$

Now we say that some test instance belongs to the target class or lies inside the hyper-sphere if $F(z) \geq 0$. Otherwise, it is an outlier lying outside the hyper-sphere. Similarly, in traditional support vector machines and other kernel machines, the inner product between two vectors in 7 and 8 can be replaced by various kernels satisfying the Mercer theorem [8].

As we see, the methodology above gives efficiently an optimized domain description particularly when there is no confrontation with evolutionary data. Nevertheless, the challenge becomes valuable when dealing with non stationary data likely to progress in the projection space. In the next section, we address this paradigm.

III. NON-STATIONARY SVDD

A. Initial formulation

The main challenge is how to deal with the newly added training samples and on which criteria we rely to discard irrelevant data. At each iteration, the optimization process is performed on training set denoted $S_{training}$. As initialization, let $S_{training} = SVs$. As previously mentioned, two main aspects will be addressed; the incremental learning and the dynamic adaptation. The adopted approach is structured around basic proprieties that we analyze in the sequel.

Propriety.1 : In the incremental learning scheme at each step we add one instance to the training set. If the newly added sample meets the KKT conditions, it has not any effect on the previous data description and can be discarded from the training set. Therefore, the optimization process is useless in this case since the new added sample doesn't minimize the currently minimum objective function and the same result will be produced. This property leads us to reduce considerably the cardinality of training samples and ensure the rapidity of convergence.

Propriety.2 : On the other hand, if samples in the newly added training set lie outside the hyper-sphere, some among them will become new support vector surely. This propriety indicates that in the newly added training set the samples violating the above KKT conditions should be taken and integrated within the training working set during the optimization process.

Accordingly, as a first step towards an incremental learning, we should seek the neighborhood of each instance presented online with respect to the current data domain. To do so, for a random sample $z^{(k)}$ coming at iteration k , the distance to the center of the hyper-sphere can be calculated through formula 8. Thus, two cases are possible here:

Case.1 : $F(z^{(k)}) \geq 0$

This case means that the query sample $z^{(k)}$ lies inside the hyper-sphere (target). According to *propriety.1*, this sample has no effect on the current SVs set and don't minimize the objective function. Hence, we preserve the same whole of SVs and we don't carry out any change on it. In this way, we get rid for once of the optimization process which can be expensive in terms of time and memory while guaranteeing the optimal solution.

Case.2 : $F(z^{(k)}) < 0$

Each sample satisfying this condition is considered as outlier and likely to be probably a new support vector according to *propriety.2*. In this case, the SVs set needs to be updated to incorporate samples newly detected. To improve performances, we assign a dynamic aspect to the learning process. Each insertion of a new sample must be accompanied by a deletion procedure applied on irrelevant data. We aim thus a removal/insertion procedure.

B. Dynamic data removal /insertion

1) *Removal procedure*: The removal procedure, as well the insertion procedure, is performed only when a random sample $z^{(k)}$ received at iteration k is an outlier meeting the condition mentioned in case.2. Otherwise, $z^{(k)}$ is assigned to $NSVs$ set. The deletion concerns particularly the irrelevant samples that don't follow the evolution of data domain. F. Camci proposed in [9] a weighted support vector novelty detector (*WSVND*) for non-stationary data. To explicitly support the non-stationary nature of the data, the method incorporates the notion of weight or importance of a data point based on its age. It forces the support vectors to be as young as possible so as to be able to effectively track the non-stationary process. The solution seems biased and insufficient given that we can fall into the case where old support vectors can be more significant than younger some others. In this way, the method is not immune to local optima drawbacks if support vectors are judged by their oldness. In view of that, we propose a simple formulation to deal with this phenomenon. Instead of judging data with respect to their oldness, we perform a weighted relevance measure based on the neighborhood concept. Indeed, samples lose their significance while moving away from the high-density zone of the data domain. Based on this assumption, a spatial trend prediction needs thus to be realized. As a notation, for each coming $z^{(k)}$, let $a^{(k)}$ and $a^{(k+1)}$ be respectively the current center of the hyper-sphere and the probable future center when $z^{(k)}$ becomes a new support vector. Let also $d^{(k)}$ and $d^{(k+1)}$ be the distance between a sample $x_i \in SVs$ and respectively $a^{(k)}$ and $a^{(k+1)}$. We attempt to extract the element among the whole SVs to be deleted. This removable support vector SV_{del} should satisfy

the following expression:

$$SV_{del} = \arg \max \left(\text{sign}(\Delta d) \times \left\| z^{(k)} - x_i \right\| \right), \quad (10)$$

$$x_i \in SVs, i = 1, \dots, N_{SVs}$$

The term Δd is the distance variation expressed as:

$$\Delta d = d^{(k+1)} - d^{(k)}, \quad (11)$$

where

$$\begin{cases} d^{(k)} &= F(x_i), \quad x_i \in SVs, \quad i = 1, \dots, N_{SVs} \\ d^{(k+1)} &= \left\| x_i - a^{(k+1)} \right\| \end{cases}. \quad (12)$$

The predicted center $a^{(k+1)}$ can be simply determined by a recursive relation, so that :

$$a^{(k+1)} = a^{(k)} + \frac{1}{N_{SVs} + 1} \left(z^{(k)} - a^{(k)} \right) \quad (13)$$

According to expression (10), the rule chooses for removal the foremost support vector which moves away from center when the last one shifts relative to each sample newly appended. This is seems proper as much as the removable support vector becomes increasingly isolated and far-off from the high-density zone of the data domain. Once SV_{del} is determined, the training set $S_{training}$ is afterwards adjusted by deleting SV_{del} and inserting the new coming instance $z^{(k)}$, so that:

$$S_{training} = \{S_{training}\} \setminus \{SV_{del}\} \cup \{z^{(k)}\} \quad (14)$$

Evidently, the instance $z^{(k)}$ must satisfy the case.2.

2) *Insertion procedure*: At this stage, to enrich the training set and guarantee an optimal description of the data domain, we aim to find data points among the $NSVs$ set which can be probably new support vectors. Without doubt, these points are those in close proximity to the interior boundary of the hyper-sphere. For each point $x_i \in NSVs$ ($i = 1, \dots, N_{NSVs}$), N_{NSVs} is the total number of samples belonging to $NSVs$ set, samples which are most likely to be support vectors can be formulated as:

$$\sqrt{F(x_i)} \geq T, i = 1, \dots, N_{NSVs}. \quad (15)$$

The term T is a decision threshold which is fixed with respect to the data distribution estimated on the target class whole. We denote by $PSVs$ the set grouping the Probable Support Vectors meeting the conditions (15). Hence, the training set $S_{training}$ is adjusted again to insert the $PSVs$ components, so that:

$$S_{training} = \{S_{training}\} \cup \{PSVs\} \quad (16)$$

As a result, we obtain at the end a compact training set that reduces significantly the complexity of the QP problem in terms of cardinality of data and convergence time. This makes the method useful for applications requiring fast data tracking, or even online processing. Moreover, despite their

small cardinality, the training set elements are significant and well chosen through selective rules to ensure the optimality in the solution. In the sequel we give the global algorithm in pseudo-code.

Algorithm : *Non – stationary SVDD*

Input Training set $S_{training} = SVs$, a sample $z^{(k)}$

Step.1 Seek the location of $z^{(k)}$
 if $z^{(k)}$ is a target ($F(z^{(k)}) \geq 0$), then

- 1: Append $z^{(k)}$ to $NSVs$ set
- 2: Repeat Step1 to treat the next sample $z^{(k+1)}$

 else

- 3: Execute Step2

 end

Step.2 Execute the deletion/insertion procedure

- 4: Remove the sample of SVs set satisfying (12)
- 5: Append $z^{(k)}$ to the training set $S_{training}$
- 6: Insert into $S_{training}$ all samples satisfying (15)
- 7: Train the SVDD on $S_{training}$ to get a new SVs

Output SVs set giving the class boundaries

In the following, we present experimental evaluation of the proposed approach on artificial basis in order to make advantage of its efficiency against evolutionary data.

IV. EXPERIMENTS

Several classification problems need specific tools that deal with evolutionary data over the time and space. In this paradigm, we proposed above a new method based on an adapted SVDD. In order to evaluate its performances, some experiments are conducted on two different datasets. The first one consists of 2-dimensional Gaussian distribution generated empirically and evolving in its projection space. The second dataset is an evolutionary banana shaped distribution projected as well into 2-dimension space. For the data domain description, we employ in all cases the Gaussian Kernel function:

$$K(x_i, x_j) = \exp\left(\frac{-\|x_i - x_j\|^2}{2\sigma}\right), \forall i \neq j \quad (17)$$

The shape of the enclosing contours in data space is governed by two parameters: the scale parameter of the Gaussian Kernel σ and C the soft margin constant (regularization parameter). In table.I, we give the initialization and proprieties of the different parameters for each dataset.

In figure.1, we estimate the non-stationary SVDD boundary for the Gaussian dataset. Indeed, two main dynamic phenomena are presented here, which are the online manifestation of

Table I
INITIALIZATION AND PROPRIETIES OF EXPERIMENTAL DATASETS.

	Distribution	Non-stationary	Kernel function
Dataset#1	Gaussian	Yes	Gaussian
Dataset#2	Banana	Yes	Gaussian

	(C, σ)	Decision threshold T	Space dimension
Dataset#1	(0.1, 0.7)	σ	2
Dataset#2	(0.1, 5)	σ	2

samples and the non-stationarity of the data shape. Towards these phenomena, the classifier is dynamically updated according to the drifting distribution as the shape evolves over the time. This is performed via an insertion procedure which aims to incorporate new information, and a deletion procedure to remove irrelevant data.

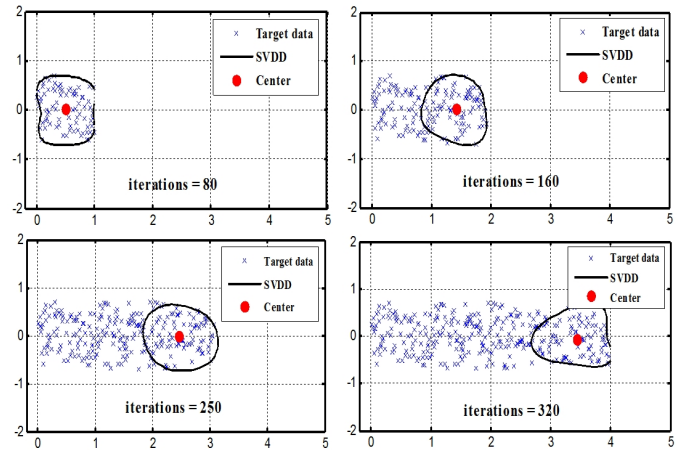


Figure 1. Novel detection for drifting Gaussian distribution.

Similarly, we apply the proposed algorithm on evolving banana dataset (figure.2). Despite dynamic phenomena previously mentioned, the data is distributed along the banana shape and superimposed with a normal distribution with standard deviation in all directions. Accordingly, the algorithm tracks properly the evolutionary shapes thanks to efficient tools leading to deal with the concept drift in non-stationary environments.

As well as the concept drift, there are still difficulties associated with SVDD applications which hamper its effectiveness and general acceptability in engineering domain. These difficulties are reflected on the two main aspects. The first one is related to the computing complexity which grows with respect to the training set enhanced at each added sample. This drawback appears when dealing with large dataset since the algorithm requires solving a quadratic programming (QP) problem in a number of coefficients equal to the number of training examples. The second difficulty is that almost all SVDD algorithms at hand are not applicable on-line, that is, in cases where data are sequentially obtained

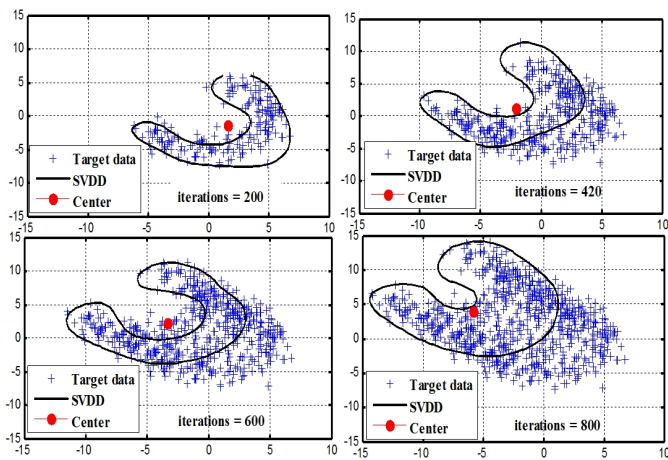


Figure 2. Novel detection for drifting Banana distribution.

and learning has to be done from the first data. These dilemmas are efficiently overcome in the present work in so far as the algorithm is endowed with useful tool by inserting the new sample and removing the most irrelevant one according to adequate rules. It acts obviously as a selective sliding window minimizing the size of the training set and avoiding the useless accrual of information. In figure.3, we illustrate the variation of cardinality of the training set $S_{training}$ and the support vectors set SVs for the two test basis (Gaussian and banana basis). As we see, the cardinality in both sets does not grow iteratively, but it fluctuates with respect to intrinsic characteristics of the distribution (density, dispersion...). In this way, we guarantee the accuracy while reducing space and time complexities.

V. CONCLUSION

In this paper, we used the SVDD technique as a robust tool for novel detection and discerning between target class and outliers. We considered the situation in which the decision boundaries are subject to concept drift and training samples are not fully available, but they are obtained sequentially one by one. We proposed a methodology for tracking evolutionary shapes and the dynamic maintaining of SVDD. This methodology is based on iterative procedure inserting each new added sample within the training set and removing useless data through adequate criteria. In order to reduce the computational complexity, based on KKT conditions, the training set is considerably reduced so that only significant data are taken to achieve the optimization process. The method is assessed thereafter on some artificial basis to prove its effectiveness.

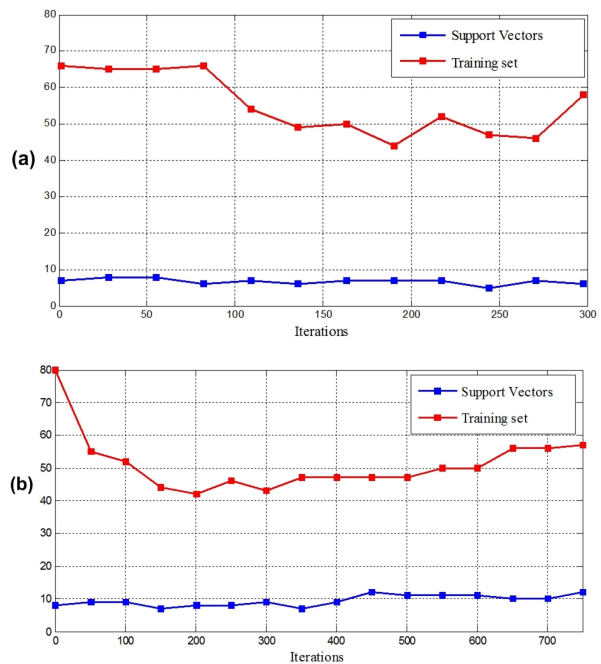


Figure 3. (a) Cardinality variation of the training set and the support vectors set associated to the Gaussian distribution. (b) Cardinality variation of the training set and the support vectors set associated to the banana distribution.

REFERENCES

- [1] D. Tax and R. Duin, *Support vector domain description*, Pattern Recognition Letters, Vol. 20, pp. 1191-1199, November 1999.
- [2] D. Tax and R. Duin, *Support vector data description*, Machine Learning, Vol. 54, pp. 45-66, January 2004.
- [3] X. Hua and S. Ding, *Incremental Learning Algorithm for Support Vector Data Description*, Journal of Software, Vol. 6, No. 7, pp. 1166-1173, July 2011.
- [4] R. R. Sillito and R. B. Fisher, *Incremental one-class learning with bounded computational complexity*, Proceedings of the 17th international conference on artificial neural networks (ICANN'07), Berlin 2007.
- [5] A. Tavakkoli, M. Nicolescu, M. Nicolescu, G. Bebis, *Incremental SVDD Training: Improving Efficiency of Background Modeling in Videos*, Proceedings of the International Conference on Signal and Image Processing, Kailua-Kona, Hawaii, August 2008.
- [6] E. Osuna, R. Freund, F. Girosi, *Improved Training Algorithm for Support Vector Machines*, Proceeding of Neural Networks in Signal Processing (IEEE NNSP'97), Amelia Island, September 1997.
- [7] X. Jianhua, *Constructing a Fast Algorithm for Multi-label Classification with Support Vector Data Description*, Proceedings of the 2010 IEEE International Conference on Granular Computing (GRC '10), IEEE Computer Society Washington, DC, USA 2010.
- [8] V. N. Vapnik, *Statistical Learning Theory*, New York: John Wiley & Sons, 1998.
- [9] F. Camci and R. B. Chinnam, *General support vector representation machine for one-class classification of non-stationary classes*, Pattern Recognition, Vol. 41, Issue 10, pp. 3021-3034, October 2008.

The Usage of Malay Technological Terminologies in Malaysian Youth Institutions for Skills: Interests and Challenges

Adenan Ayob
Sultan Idris Education University
35900 Tanjung Malim
Perak, MALAYSIA
adenan@fbk.upsi.edu.my

Abstract - A very common problem of terminology work is the importance and indeed the very nature of terminology that poorly understood. Thus many people simply have no idea at all of what it is, while others searching for an explanation of some sort, end up associating it with different meaning. Related professions in the communications field, such as translation and technical will often be aware of the word without having precise knowledge of what it entails. This paper highlights the interests and challenges of the usage of Malay technological terminologies in Malaysian youth institutions for skills. Other than discussing the meaning of terminology and technology and their concept, this paper recommends the proposed methodology for the development of technology terminologies that based on the theory of Post-Modernism and Constructivism. That theory has made coverage of scientific research for International Energy Agency. The main aims under those discussions were to discover and to ensure that the terminologies to be developed to function automatically and to accelerate the collection process, the creation and dissemination.

Keywords: Malay technological terminology, skills, interests, challenges

I. INTRODUCTION

Before pioneering the interests and challenges, the main problem is related to the area of using foreign terminologies, especially the terminologies that borrowed and modified from other languages. Notably, the initiative to introduce new terminologies in Malay language technology has been implemented for a long time since the establishment of Malaysia National University in 1970. The implementation of technological programs in Malaysia National University covers many important aspects of terminology.

The fact that is desired, the Malaysian youth institutions for skills can refer to various resources that based on the technological terminologies from dictionary and glossary based on information and communication technology as published by the Malaysian government. The portal that organized by the government too is a popular reference at this time to ensure that the terminologies used are accurate. However, the actual reference material produced specifically for Malaysian youth institutions for skills should be issued regarding certain technological courses.

With a variety of platforms that can be accessed by each student at the institute, either through internet or printed materials, the issue of an inappropriate use of terminologies will not arise again. Steps taken through creation of technological terminologies were to cover a lot of problems on the widespread of foreign terminologies. This gives the impression that the usage of technological terminologies is more important.

There's no denying that the use of more precise technological terminologies in those institutes may be awkward for some people. However, the responsibility of accurate terminology is the role of all parties. The management and students of those institutions should be more sensitive to the development of terminology. In fact, the reference could be easily accessed, especially with the versatility of information and communication technology, computers and smart phones available.

Many benefits can be achieved through technology equipment. Area of internet access has also been established in Malaysia. Thus, when faced with a situation to choose the right terminologies, students can easily reach relevant portal for reference. Other alternative might be through dictionary and glossary references. Dictionary and glossary, reference should be

made in the selection of an appropriate and accurate terminology.

II. TECHNOLOGY

Technology, according to Hervey and Higgins (1992) is closely linked with science and engineering. In other words, the technology consists of two dimensions, namely science and engineering that are interconnected with each other. Science refers to understanding the real world in the human environment. This means that technology is a basic feature in the dimensions of materials and power and the interaction of science.

In fact, according to Brislin and Richard (1976), science is a medium that ultimately create a culture and an expression of physics. According Brislin and Richard too, the engineering is a matter connected with knowledge about materials and power that applied in the areas of planning, including technical matters. In another sense, technology accounts for techniques and tools for maintaining a work that based on the results of science.

III. TECHNOLOGICAL TERMINOLOGIES CONCEPTS

Technology by Catford (1965) conceptually have three principles, namely (i) technology; human artifacts, including the hardware and the system of large-scale complex technology, (ii) the nature of technological creation and discovery, development and dissemination to the public widely, and (iii) technology, which began from a very specific technique and practically involving scientific technology systems. Therefore, technology is defined as the study of the relationship between human and the world, which manifests itself in view of technology, research on the phenomenon of the overall technology, placement of technology in community development retrospectively and prospectively in accordance with those dimensions.

According to technology research purposes by Brislin and Richard (1976), it is focused on the technical sciences or engineering, technical products, activities and knowledge as a cultural phenomenon. Brislin and Richard added that technology involves technical development and knowledge.

Similar views by Nida, Eugene A (1964), it is also related to technological progress and its relationship with the philosophy of technology. Nida, Eugene A divides the philosophy of technology in two stages, namely the cognitive and instrumental. Each stage will be followed by rapid technology change and vary.

IV. THE IMPORTANCE OF USING MALAY TECHNOLOGICAL TERMINOLOGIES IN MALAYSIAN YOUTH INSTITUTIONS FOR SKILLS

The technological terminology in Malaysian youth institutions for skill was centered on the importance of an underlying scientific basis and fundamental science. These interests include applied science and applied research. Actual sustainability terminologies should be seen as two bands that are complementary and generate the ideas that based on tools and materials.

In cognitive aspect, knowledge that based on technology is designed efficiently to resolve practical problems (Arrowsmith & R. Shattuck, 1961). Technology changes the capacity to apply scientific aspects in the development of technological knowledge.

In instrumental level, technology is a set of artifacts that are designed and produced intensive to perform the functions of mechanical and electronic (Arrowsmith & R. Shattuck, 1961). Changes in the instrumental-based technology is consistent and in line with the aspirations of the development of technology. The underlying technology of the entity is not a system of knowledge, but rather a complex system designed from the intentional operator.

Technology is the medium of terminology resources that easily and quickly manage to acquire knowledge. Awareness of the importance of terminology has prompted adoption of skills needed by students to compete in the new millennium.

Literacy in terminologies of learning technologies can purpose efficiently and effectively. So, rapid progress in technology allows a student to collect, transmit, distribute, manage, process or store various types of information quickly and easily (Bell, 1911). Technology is considered a new facility in the smart education system.

The usage of the terminologies in youth institution for skill is the first step towards creating a technological society in line with the progress of the country. The terminologies that controlled technology allowed students to easily master the knowledge and skills in order to use the facility for the challenges in industrial world.

V. THE CHALLENGES OF USING MALAY TECHNOLOGICAL TERMINOLOGIES IN MALAYSIA YOUTH INSTITUTION FOR SKILLS

Among those challenges, the terminologies that exist in communications felt awkward when attempted to be preserved. This was probably due to an inappropriate interpretation of such terminologies. Such complexity also gives effect to the interpretation. Interpretation involves a source language in the recipient. In this process, it should be identified by the expression equation method based on the meaning of the source language and style or the way of common language. The aim is to improve the ability of consumers to make timely and accurate interpretation based on the formation of a creative and innovative thinking.

The above statement reflects the existence of problems in any absolute balance. An axis of problems often emerged from the resolution of terminologies due to constraints of remuneration, strings, and the coefficient in plural noun. Therefore, the importance of interpretation for such terminology has not been formulated as a whole. In fact, to show any strong interpretation of terminologies should be consistent with a principle or theory.

This means that there is no expression on the interpretation of terminologies to suit a model equation of language sources. In other sense, all kinds of interpretations of terminologies may be an addition and distortion of information.

Clearly, in order to interpret accurately and neatly on the barrel of technical terminologies, a complex problem will always exist. Thus, this study will focus on a suitable method that is characterized by the terminology development theory based, including the interpretation under practical facilities. The terms may not be able to carry the intended meaning. Therefore, the interpretation of terminologies should be attributed to the expression of needs-based communication process.

The study to be conducted was focused on the attitude that based on the dimensions of (i) perceptual; knowledge and confidence, (ii) affective; facilitate and (iii) behavior. The practical dimensions are involved (i) strategies, (ii) the level and (iii) interest, while a poll among experts will focus on the content and appropriateness of the terminology which is based on the curriculum and syllabus, and the importance of practice and interpretation procedures, including practicalities study that based upon a degree course.

VI. PROPOSED METHODOLOGY

Recommendations proposed methodology for the development of technology terminologies is based on the theory of Post-Modernism and Constructivism. That theory has made coverage of scientific research for International Energy Agency which is adapted to the theory of Post-Modernism and Constructivism. In the countries concerned with several components of the source terminologies development in technology, students tend toward the formation of attitudes and practices of the sustainability of a technical course. Therefore, it can be justified that the youth in Malaysia should realize the vision for the interpretation of terminologies might think the importance of technology in building strong technical thinking.

Theory is implicit in the concept of sustainability accuracy of an interpretation towards terminology. According to The Partnership for a Combination of 21st Century Skills or P21, the accuracy of an interpretation is a concept of sustainability in the appearance of a technical knowledge that to be mastered. This includes the ability for a student to master the strategies and skills related to creative and innovative thinking, as well as other skills across towards establishing autonomy in thinking. Intended interpretation of the concept of accuracy is not limited to the control of terminologies, but also the feasibility and application of technical skills to meet all demands for the development of an industry. The command interpretation for the 21st century youth should give the ability to mutually combine local knowledge with global knowledge and make them flexible and efficient to adapt the form of knowledge and also sharing knowledge that is constantly changing and do not marginalized or left behind.

VII. CONCLUSION

Under the technological terminology, the work of lexicography should be seen as an important role in generating balance language and technology development. To compete with other foreign terminologies, a development of more innovative methods deemed appropriate. This effort is at least in the context of terminologies collection that to be implemented. The proposed methodology involves the collection of texts relating to specific areas of the experimental identification of terminologies under the supervision of local experts and related to the perform analysis of such courses.

Therefore, the usage of the terminologies that regards upon research at these institutions should be seen as an early step in the construction of technological system. The aim is to ensure that the terminologies to be developed to function automatically and to accelerate the collection process, the creation and dissemination.

ACKNOWLEDGEMENT

This research is my original work. The research contributes to the acquisition of terminologies and technological concepts.

REFERENCES

- [1] Arrowsmith, W., and R. Shattuck, *The craft and context of translation*. The University of Texas Press, 1961.
- [2] Bell, R.T., *Translation and translating: Theory and practice*. London: Longman, 1911.
- [3] Brislin, Richard W. (ed.), *Translation: Application and research*. New York: Gardner Press, 1976.
- [4] Catford, J.C., *A linguistics theory of translation*. London: Oxford University Press, 1965.
- [5] Hervey, S. Higgins, I., *Thinking translation*. London: Routledge, 1992.
- [6] Nida, Eugene A., *Toward a science of translation*. Leiden: E.J. Brill, 1964.

Analytical Solution of an Electrokinetic Flow in a Nano-Channel with Variable Physical Properties

Mehdi Mostofi

Department of Mechanical Engineering
East Tehran Branch, Islamic Azad University
Tehran, Iran
mehdi.mostofi@gmail.com

Abstract—In this paper, an electrokinetic flow of a water electrolyte in a nano-channel will be studied. This study will be with existence of the Electric Double Layer (EDL) and fully analytical. Governing equation for the EDL is Poisson-Boltzmann. In addition, Navier-Stokes equations for electrolyte flow, Species and mass conservation equations are in use. Induced electric potential force the electrolyte ions and decrease the mass flow rate. In this paper, effect of temperature rise on the physical properties and consequently, on the velocity and potential distribution will be investigated.

Keywords- *Electrokinetics, Electric Double Layer (EDL), Zeta Potential, Nano-Channel, Variable Physical Properties, Temperature Variations.*

I. INTRODUCTION

In recent decades, after introducing micro- and nano-fabrication technologies, several possibilities in the case of micro- and nano-fluidic devices have been invented. This idea has been followed by some modern technologies such as Lab-on-a-Chip.

One of the most important subsystems of the micro- and nano-fluidic devices is their passage or “Micro- and nano-channel”. Nano-channel term is referred to channels with hydraulic diameter below 100 nm. [1]. By decrease in size and hydraulic diameter some of the physical parameters such as surface tension will be more significant while they are negligible in normal sizes.

Concentrating surface loads in liquid – solid interface makes the EDL to be existed. If the loads are concentrated in the end of nano-channels, a potential difference will be generated that forces the ions in the nano-channel. However, induced electric field is discharged by electric conduction of the electrolyte.

The first significant work that was done in the literature belongs to 1870 that Helmholtz introduced the EDL. According to this finding, flow and electricity parameters for electrokinetic transport were detected. Electroosmotic processes have been utilized since 1930s. Modern theoretical progresses in the case of electrokinetic flow can be found in [2 – 6]. Burgreen and Nakache [2] and Oshima and Kondo [3] studied the flow between two parallel plates. Also, Rice and Whitehead [4], Lo and Chan [5] and Ke and Liu [6] studied the

flow in capillary tube. Solving the problem considering this fact is necessary. In this paper, for small zeta potentials without pressure gradient will be studied based on the curvilinear coordinates in a capillary tube. In continue, physical properties will be assumed to be variable in order to investigate the effect of temperature variation on potential and velocity distribution.

II. METHODS

First, In electrokinetic processes, for the most general form of the study, four types of equations are used [7]:

- Conservation of species (same as number of ion species can be found in an electrolyte), at least 2 equations.
- Conservation of mass, 1 equation.
- Poisson-Boltzmann equation, 1 equation.
- Navier-Stokes equations, 3 equations.

It is clear that, at least seven nonlinear equations govern an electrokinetic process. In this paper, by some simplifications that will be mentioned later, this set will be made simpler as follows:

$$\frac{1}{r} \frac{\partial}{\partial r} \left(r \frac{\partial \varphi}{\partial r} \right) = \frac{-\eta}{\varepsilon^2} (X_p - X_n) \quad (1)$$

$$\frac{1}{r} \frac{\partial}{\partial r} \left(r \frac{\partial u}{\partial r} \right) = \frac{-\eta \varepsilon_e E_0 R T}{\varepsilon^2 F \mu U_0} (X_p - X_n) \quad (2)$$

$$\frac{1}{r} \frac{\partial}{\partial r} \left[r \left(\frac{\partial X_p}{\partial r} + \xi X_p \frac{\partial \varphi}{\partial r} \right) \right] = 0 \quad (3)$$

$$\frac{1}{r} \frac{\partial}{\partial r} \left[r \left(\frac{\partial X_n}{\partial r} - \xi X_n \frac{\partial \varphi}{\partial r} \right) \right] = 0 \quad (4)$$

Where in these equations, r refers to dimensionless radial position that is scaled by tube radius, φ is dimensionless potential that is scaled by zeta potential φ_0 , X_p is cation specie concentration (in this paper Na^+), X_n is anion specie concentration (in this paper Cl^-), u is dimensionless velocity that is scaled by scale velocity U_0 , η is ionic strength, ε is

Debye-Huckel parameter that is the ratio of EDL thickness to tube length, ϵ_e is dielectric constant of the fluid (in this paper, water), E_0 is electric energy strength (V/m), R is global gas constant, T is temperature, F is Faraday Constant and μ is fluid viscosity.

In this paper, we discuss about the problems with small zeta potentials. According to Boltzmann distribution, we can derive [8]:

$$X_i = \exp\left(\frac{-z_i e \varphi^*}{k_b T}\right) \quad (5)$$

Which z_i is valence of the species in electrolyte, e is charge of the electron, φ^* is potential (V) and k_b is Boltzmann constant. On the other hand, for positive ion (subscript p), $z = 1$ and for negative one (subscript m), it is -1 . As a result, (1) can be rewritten as:

$$\frac{1}{r} \frac{\partial}{\partial r} \left(r \frac{\partial \varphi}{\partial r} \right) = \frac{1}{\epsilon^2} \sinh(\varphi) \quad (6)$$

Which in this equation:

$$\varphi = \frac{\varphi^* e}{k_b T} \quad (7)$$

A. Modeling and Simulation of Potential Distribution

In this paper, it is assumed that, zeta potential is so small that, we can consider:

$$\sinh(\varphi) = \varphi \quad (8)$$

As a result, (6) can be rewritten as:

$$\frac{1}{r} \frac{\partial}{\partial r} \left(r \frac{\partial \varphi}{\partial r} \right) = \frac{\varphi}{\epsilon^2} \quad (9)$$

On the other hand:

$$r^2 \varphi'' + r \varphi' - \frac{r^2}{\epsilon^2} \varphi = 0 \quad (10)$$

As it can be seen, the latest equation is a simple form of the Developed Bessel ODE [9]. Based on the instructions in [9], if we have:

$$r^2 \varphi'' + r(a + 2bx^p) \varphi' + [c + d r^{2q} + b(a + p - 1)r^p + b^2 r^{2p}] \varphi = 0 \quad (11)$$

B. Simulations with Variable Physical Properties

In this paper, we investigate the above equations by using variable physical properties. Fig. 1 shows the variations of viscosity relative to temperature [10].

In addition, variations of dielectric constant over temperature and the effects of these variations have been investigated. Fig. 2 shows the variations of dielectric constant over temperature [11]. It is assumed that, all phenomena are in standard pressure ($p = 100$ kPa) and it is constant.

According to the above assumptions, we can investigate the effect of temperature variation on the potential and velocity profiles. Fig. 3 shows the temperature variation effect on the zeta potential on the tube wall. In this figure, it is assumed that, the potential quantity in $T = 27^\circ\text{C}$ is 1. It is clear that, the potential distribution is linearly dependent of the temperature.

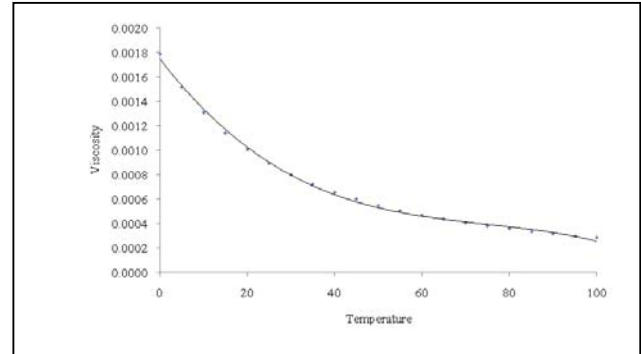


Figure 1. Variations of viscosity in relation to temperature

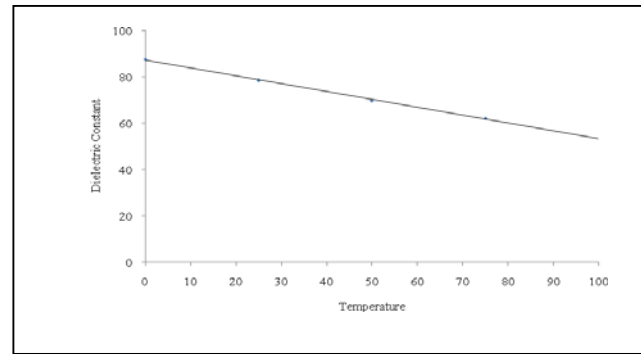


Figure 2. Variations of water dielectric constant as a function of temperature ($p = 100$ kPa)

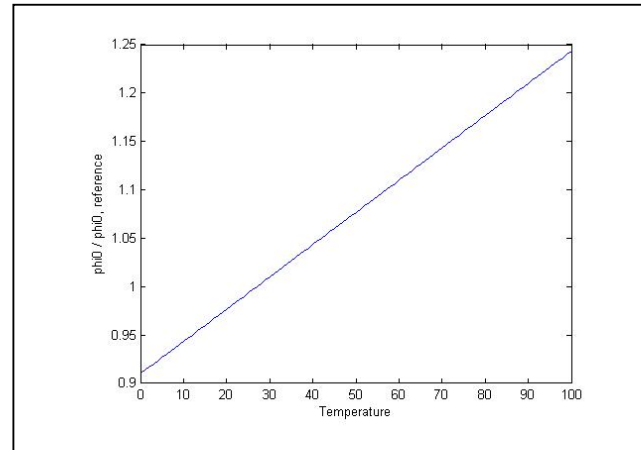


Figure 3. Temperature variation effect on potential distribution (vertical quantity is 1 for $T = 300$ K)

According to the fact that, velocity is dependent of zeta potential, dielectric constant and viscosity, velocity can be differed by the temperature variation. Because of the fact that, temperature variations have different effects on the mentioned quantities, Fig. 4 can be generated. As it can be seen, the variations are increasing but not uniform. Also in this case, it is assumed that, velocity scale in $T = 27^{\circ}\text{C}$ is the reference value.

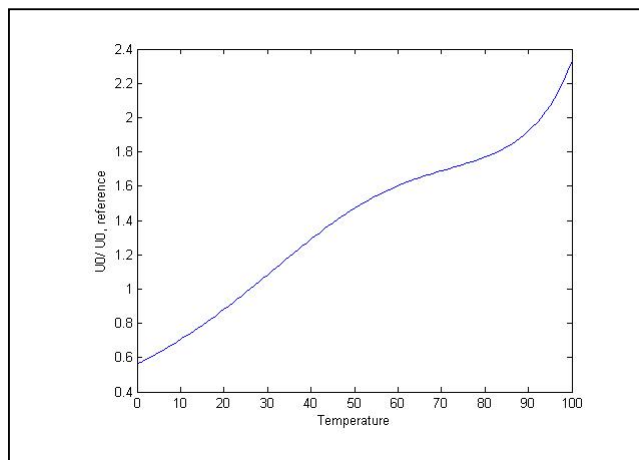


Figure 4. Temperature variation effect on velocity distribution (vertical quantity is 1 for $T = 300\text{ K}$)

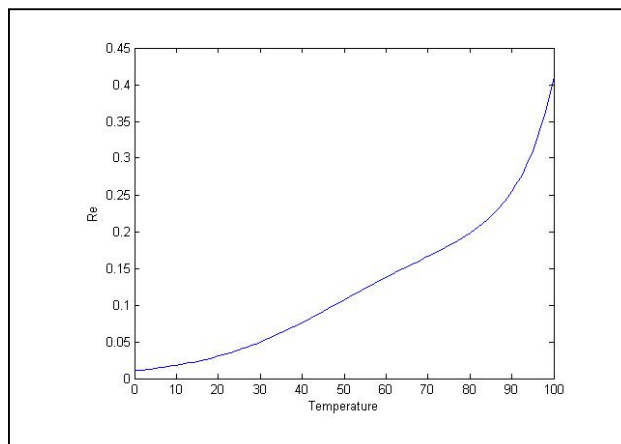


Figure 5. Temperature variation effect on Reynolds number (variable reference velocity, variable viscosity and 30 nm diameter)

Furthermore, Fig. 5 shows the temperature effect on Reynolds number in scale velocity. In this figure, it can be validated why we just consider the viscous term in Navier-Stokes equations. In this case, diameter of the tube is 30 nm.

III. RESULTS AND DISCUSSION

Before In this paper, velocity and potential distribution in four specific temperatures ($T = 20, 40, 60$ and 80°C) has been investigated. Figs. 6 and 7 show the potential and velocity distribution over the tube radial position in these four specific temperatures. As it can be seen in Figure 8, temperature rise has a small increase in potential distribution. It is notable that,

if temperature rises from 20°C to 80°C , peak potential has at most 21% increase. Also, rates of decrease from zeta potential to zero are not significantly different in different temperatures.

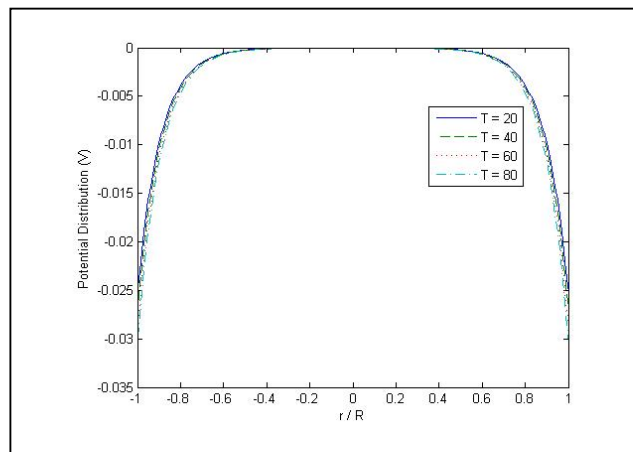


Figure 6. Potential distribution over nano-tube in different temperatures

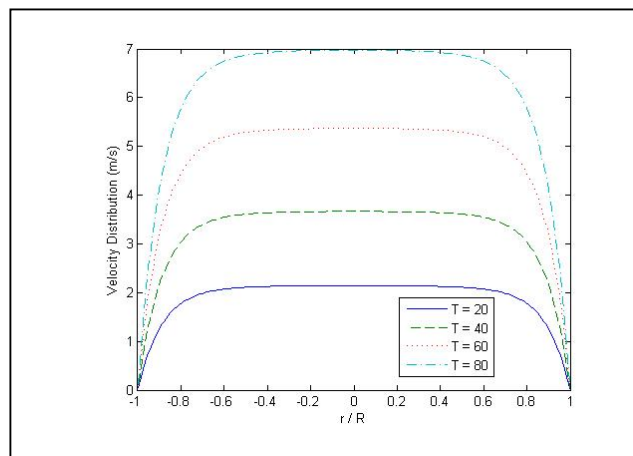


Figure 7. Velocity distribution over nano-tube in different temperatures

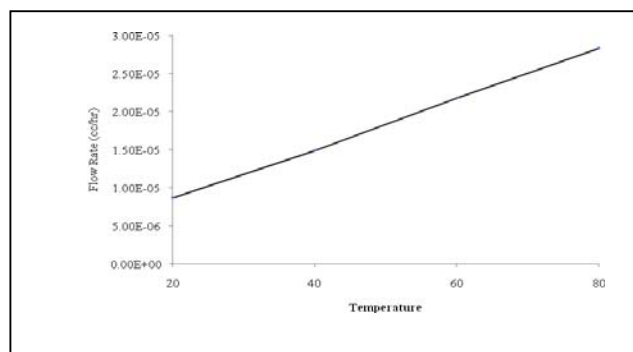


Figure 8. Flow rate in different temperatures

Considering Fig. 7 notes that temperature rise from 20°C to 80°C results in velocity rise, much more significant rather than potential distribution. In this case, if temperature rises from 20°C to 80°C , maximum velocity will rise in about 326%. Of

course, flow rate increase by temperature rise, is not as significant as of maximum velocity in nano-tube (Fig. 8). In addition, rates of velocity increase from zero to scale university are different in different temperatures.

IV. CONCLUSION

In this paper, an analytical solution for electrokinetic flow in a nano-tube with small amount of zeta potential has been done. In this case, by considering curvilinear coordinates and using Taylor series, some derivation of Developed Bessel ODE has been derived and solved for Poisson-Boltzmann equation. In addition, velocity profile in nano-tube has been achieved. In addition, by assumption of variable physical properties such as zeta potential, dielectric constant and fluid viscosity according to temperature variation, effect on potential distribution and velocity distribution have been investigated. In both cases, it has been found out that, temperature rise will result in increasing zeta potential and velocity scale. It has the most significant effect on the velocity distribution. Simulations showed that, if we increase the temperature from 20 to 80°C, scale velocity will rise 326% and volumetric flow rate will rise significantly as well.

REFERENCES

The template will number citations consecutively within brackets [1]. The sentence punctuation follows the bracket [2]. Refer simply to the reference number, as in [3]—do not use “Ref. [3]” or “reference [3]” except at the beginning of a sentence: “Reference [3] was the first . . .”

Number footnotes separately in superscripts. Place the actual footnote at the bottom of the column in which it was cited. Do not put footnotes in the reference list. Use letters for table footnotes.

Unless there are six authors or more give all authors' names; do not use “et al.”. Papers that have not been published, even if they have been submitted for publication, should be cited as “unpublished” [4]. Papers that have been accepted for publication should be cited as “in press” [5]. Capitalize only the first word in a paper title, except for proper nouns and element symbols.

For papers published in translation journals, please give the English citation first, followed by the original foreign-language citation [6].

- [1] Kandlikar S.; Garimella S.; Li, D.; Colin S.; and King M.R. Heat Transfer and Fluid Flow in Minichannels and Microchannels; Elsevier Limited: Oxford, 2006.
- [2] Burgreen D.; and Nakache F.R. Electrokinetic Flow in Ultrafine Capillary Slits. *J. Phys. Chem.*, 1964, 64(5), 1084–1091.
- [3] Oshima H.; and Kondo T. Electrokinetic Flow between Two Parallel Plates with Surface Charge Layers: Electroosmosis and Streaming Potential. *J. Colloids and Interface Science*, 1990, 135(2), 443–448.
- [4] Rice C.L.; and Whitehead R. Electrokinetic Flow in a Narrow Cylindrical Capillary. *J. Phys. Chem.*, 1965, 69(11), 4017–4023.
- [5] Lo W.Y.; and Chan K. Poisson-Boltzmann Calculations of Ions in Charged Capillaries. *J. Chem. Phys.*, 1994, 143, 339–353.
- [6] Keh H.; and Liu Y.C. Electrokinetic Flow in a Circular Capillary with a Surface Charge Layer. *J. Colloids and Interface Surfaces*, 1995, 172, 222–229.
- [7] Conlisk A.T.; McFerran, J.; Zheng, Z.; and Hansford, D. Mass Transfer and Flow in Electricity Charged Micro- and Nano- Channels, *Anal Chem.*, 2002, 74(9), 2139–2150.
- [8] Zheng, Z. Electrokinetic Flow in Micro- and Nano- Fluidic Components. PhD Thesis, Ohio State University, Ohio, USA, 2003.
- [9] Shahani, A.R. Advanced Engineering Mathematics; K.N. Toosi University of Technology Publications: Tehran, 2000.
- [10] Wylie, V.L.; and Streeter, E.B. Fluid Mechanics: First SI Metric Edition, Mc-Graw Hill, 1983.
- [11] Uematsu, M.; and Franck, E.U. Static Dielectric Constant of Water and Steam, *J. Phys. Chem. Ref. Data*, 1980, 9(4), 1291–1306.

Use of Informational Technologies in Study of the Spatial Structure of Glyprolines

L.I.Ismailova, R.M.Abbasli, S.R.Akhmedova, N.A.Akhmedov

Institute for Physical Problems, Baku State University,
Z. Khalilov Str.23, Baku, AZ-1148, Azerbaijan
E-mail: Abbasli_Rena@mail.ru

Abstract- XXI century is called information century. The role of Informational Technologies, modern computer programs in studying of the spatial structure of the peptide molecules in living systems is very important. One of the basic problems for molecular biophysics is investigating their structure–functional organization. Computer modeling helps to solve this problem. This work is devoted to study the spatial organization, conformational possibilities of the glyproline molecules for the first time. The calculations were carried out by the method of theoretical conformational analysis and a special computer program. Basing the calculation results the simulated analogues of each peptide molecule are constructed, the conformational mobility of the amino acid side chains are investigated and the amino acids with specific interplays with different receptors are founded

Key words: molecule, peptide, structure, conformation, analogue

I. INTRODUCTION

The successful development of computer technology and software allow us to solve actual problems of molecular modeling three-dimensional structure of peptide molecules. Peptides regulate all functions of a living organism. Using the regulatory peptides of the human body, you can create new and effective drugs.

It is known, that stress and inflammation lead to disorders in rats mesenteric microcirculatory system. These disorders are connected with mast cells activation. Proline containing peptides had a protective effect in microcirculatory dysfunction under conditions of inflammation and stress. This effect can be connected with these peptides ability to stabilize mast cells. Glyprolines are a new family of biologically active peptide drugs containing Gly (G) and Pro (P) amino asides in their structure [1-3]. Glyprolines are fragments of collagens. These molecules modulate the nervous and immune system, possess antiulcer action. At present, the various synthetic analogues of natural glyprolines were founded (Simax and Selank) [4-6].

Glyproline peptide family includes the simplest proline-containing linear peptides PG, GP, PGP, (GP)₃, (PG)₃. The biological functions of these peptides in living systems are related with their specific spatial structures. To understand the mechanism by which the glyprolines function, it is necessary to know their spatial structures and the full complement of low-

energy conformational states. The aim of this article is to study the structural organization of four groups of peptide molecules: hexapeptides GPGPGP, PGPGP, PGPPGP, heptapeptides Simax and Selank. The present paper is an extension of our previous investigations of structural and functional organization of peptide molecules [7-9].

II. METHOD

Calculation of glyproline peptides has been carried out by the method of theoretical conformational analysis with regard to nonvalent, electrostatic and torsional interactions and energy of the hydrogen bonds. In presenting the results of the calculation of the spatial structure of the molecules we used the classification suggested in the work [10]. According to it all structural versions break down into shapes including certain forms of the main chain and each form is represented by a set of conformations. The conformations are determined by the number of rotational degrees of freedom of the side chains of the residues being included in the molecule.

The conformational state of each amino residue is conveniently described by the backbone φ , ψ and side chain χ_1 , χ_2 ... dihedral angles. The terms “conformation” used in the following analysis will always imply exact quantitative characteristics of residue or fragment geometry. For a stable conformation, the φ and ψ dihedral angles are located in low-energy region R, B, L and P of the conformational map. We introduce the notion “form of a residue” to denote the region of its backbone dihedral angle (R, B, L and P). The conformation of the backbone forms of residue in a given amino acid sequence will specify the backbone form of a fragment. Forms belonging to a particular shape have an analogous peptide chain contour and a similar mutual arrangement of backbones and side chains. A procedure for the minimization fragments global energy was conducted by the method of conjugate gradients using the program described [11]. Designations indications of dihedral angles have been measured up to the generally accepted nomenclature [12].

III. RESULTS AND DISCUSSION

One of the basic problems for molecular biophysics is investigating of the spatial organization of the peptide molecules and their structure–functional organization. In the

represent work the conformational analysis was applied to investigation of the spatial structure and conformational possibilities of the peptides, belonging to glyproline peptides family.

Molecule Gly-Pro-Gly-Pro-Gly-Pro

Spatial structure of hexapeptide molecule Gly-Pro-Gly-Pro-Gly-Pro has been investigated in fragments. At the first stage conformational possibilities of dipeptide Gly-Pro, tripeptide Gly-Pro-Gly, tetrapeptide (Gly-Pro)₂ have been studied. Obtained calculation results of these fragments are as initial approximations for calculating three-dimensional structure of entire molecule (Gly-Pro)₃.

Dipeptide contains 27 atoms and 6 variable dihedral angles. For this fragment may be 8 forms: extended BB, BR, LB, LR and folded forms RB, RR, PR, PB. The lowest energy conformation PR has folded form of the main chain. Tripeptide molecule contains 34 atoms and 9 variable dihedral angles. For it are possible 32 forms of the main chain. The calculation showed that most low-energy is RRB form that has folded course of the main chain. For tetrapeptide fragment (Gly-Pro)₂, which includes 48 atoms and 11 variable dihedral angles, it was formed more than 200 initial approximations. As a result, only a very restricted set of low-energy conformations was isolated from a great number of analysed combinations of the tetrapeptide fragment with relative energies in a sufficiently wide interval of 0 – 4 kcal/mol. Possible structure of the (Gly-Pro)₂ under physiological conditions may be described by a set of low-energy half-folded forms of the back bone RRBR, BRRB and folded forms of the back bone RBPR, RRRR, PRRB.

At the final stage of the analysis, a calculation of the N-terminal tetrapeptide (Gly-Pro)₂ and the C-terminal dipeptide Gly-Pro enabled us to estimate the conformational properties of the hexapeptide molecule (Gly-Pro)₃. The starting conformations of this molecule were constructed from the low-energy conformations of the tetrapeptide fragment and the stable conformations of the dipeptide fragment. Thus, at the last stage a number of structures of hexapeptide molecule to be analysed amounted to 100. We carried out all of these structures by minimization over all the dihedral angles. The relative energy of the conformations of the hexapeptide molecule varied within the range 0–10.6 kJ/mol. Table 1 presents the energy distribution of the contributions in the most preferential conformations of the molecule (Gly-Pro)₃. The global conformation of this molecule ($E_{rel}=0$ kkal/mol) is BRRBPR. The contribution of the stabilizing nonvalent to this conformation is (-11,7) kJ/mol, whereas electrostatic interactions account for (- 4,1 kkal/mol) and torsion, for 1,8 kkal/mol. The main contributions of the interresidual interactions in this conformation were: dipeptide contributions (-8,6) kkal/mol, tripeptide (-5,4) kkal/mol, tetrapeptide (-4,7) kkal/mol, pentapeptide (-3,1) kJ/mol and hexapeptide (-0,7). Figures 1 represent schematically the backbone forms and positions of residues in low-energy conformation BRRBPR of the hexapeptide molecule (Gly-Pro)₃.

Then the spatial structure and conformational properties of the glyproline molecules (Pro-Gly)₃, (Pro-Gly-Pro)₂, Simax (Met-Glu-His-Phe-Pro-Gly-Pro) and Selank (Thr-Lys-Pro-Arg-Pro-Gly-Pro) have been investigated using the

TABLE I
ENERGETICAL PARAMETERS OF LOWEST-ENERGY
CONFORMATIONS OF GLYPROLINE MOLECULES

№	Conformation	E_{nb}	E_{el}	E_{tors}	E_{total}	E_{rel}
Molecule Gly-Pro-Gly-Pro-Gly-Pro						
1	BRRBPR	-11,7	-4,6	1,3	-15,0	0
2	BRRBLR	-10,1	-3,6	2,0	-11,7	3,3
3	PRRBPR	-9,9	-4,2	1,3	-12,8	2,2
4	RRRBPR	-10,0	-4,4	2,3	-11,9	3,1
Molecule Pro-Gly-Pro-Gly-Pro-Gly						
1	RPRRRR	-13,7	-4,2	1,9	-16,0	0
2	RPRRRB	-13,0	-4,0	2,1	-14,9	1,1
3	BPRRRRL	-12,0	-4,5	2,0	-14,5	1,5
4	RRRRRR	-12,9	-4,4	4,1	-13,2	2,8
Molecule Pro-Gly-Pro-Pro-Gly-Pro						
1	RPBRRR	-12,6	-7,7	1,8	-18,0	0
2	RBBRRR	-10,3	-5,8	1,5	-15,7	2,3
3	RRBRRR	-11,5	-6,6	2,2	-16,0	2,0
4	BBRRR	-10,4	-7,0	2,2	-15,2	2,8
Molecule Met-Glu-His-Phe-Pro-Gly-Pro						
1	BRRBBLR	-29,8	0,2	2,8	-26,8	0
2	BRRBBPR	-27,4	-2,6	4,1	-25,9	0,9
3	RRRBRRR	-27,1	-1,1	3,8	-24,4	2,3
4	RRRBRBB	-25,7	-0,4	3,6	-22,5	4,3
5	RRRBRPR	-23,1	-1,3	3,7	-21,4	5,4
Molecule Thr-Lys-Pro-Arg-Pro-Gly-Pro						
1	RBRBBLR	-30,6	0,3	5,8	-24,5	0
2	RBRBRRB	-28,1	3,1	4,8	-20,2	4,2
3	BBRBBLR	-25,9	1,4	3,4	-21,1	3,4
4	BBRBBLR	-24,0	0,1	3,8	-20,1	4,4
5	RBRBRRR	-23,7	2,9	2,4	-17,6	6,9

method of theoretical conformational analysis. The conformational potential energy of each molecule is given as the sum of the independent contributions of the non-valent, electrostatic, torsional interactions and hydrogen bonds. The low-energy conformations of these molecules and the values of the dihedral angles of the main and side chains are found and the energy of the intra- and inter-residue interactions is estimated. Table 1 present the energy distribution of the contributions in the most preferential conformations of these molecules. The conformational rigid and labile segments of these molecules were revealed. The low-energy conformations of the natural peptides were used as the initial structural states

to explore the conformational possibilities of the artificial analogues.

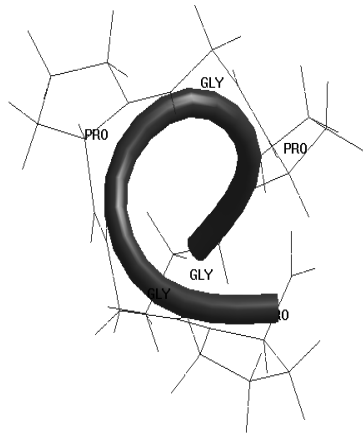


Fig. 1. Atomic model of spatial structure of (Gly-Pro)₃ molecule.

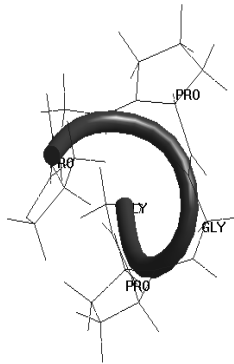


Fig. 2 Atomic model of spatial structure of (Pro-Gly)₃ molecule.

At last stage many conformational maps for side chains of amino acid residues in the low-energy conformations of the natural peptides were investigated on the basis of a semi empirical method of conformational analysis. The investigation of the molecular dynamics of each peptides is of great importance to understanding the mechanism of action these glyproline peptides with their receptors.

Glyprolines are a family of endogenous regulatory peptides, which have already shown wide range of biologic activities. The calculation of the spatial structure of peptide glyproline molecules (Gly-Pro)₃, (Pro-Gly)₃, (Pro-Gly-Pro)₂, Simax (Met-Glu-His-Phe-Pro-Gly-Pro) and Selank (Thr-Lys-Pro-Arg-Pro-Gly-Pro) found for each set of low-energy conformations of these molecules (Table 1).

The global conformations of these

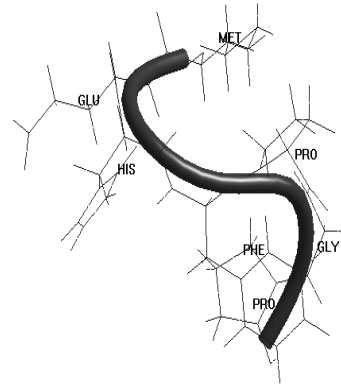


Fig. 3. Atomic model of spatial structure of Simax molecule.

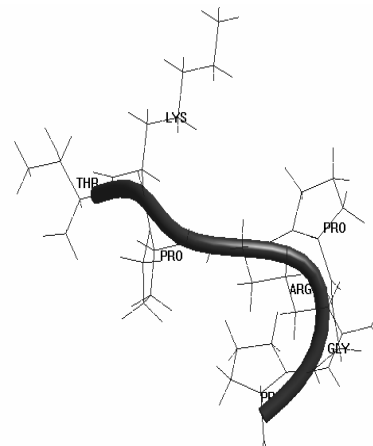


Fig. 4. Atomic model of spatial structure of Selank molecule.

molecules represent schematically the backbone forms and positions of residues in Figures 1, 2, 3 and 4. In these figures show that the C-terminal region is folded forms of the backbone of the glyproline molecules.

Comparing the results obtained, we can say that with repeated glyprolines fragments Gly-Pro and Pro-Gly-Pro essential for cytoprotective activity is the C-terminal dipeptide fragment Gly-Pro. This C-terminal dipeptide fragment Gly-Pro is present in all of the studied molecules. According to experimental data, the presence of this fragment is responsible for the protective action meet the following molecules per cell.

This work has been fulfilled in the frame of collaboration treaty of the Qafqaz and Baku State Universities.

REFERENCES

- [1] Umarova B.A., Kopylova G.N., Smirnova E.L. et al. "Secretory Activity of Mast Cell during Stress: Effect of Prolyl-Glycyl-Proline and Simax", *Bullet.of Exper. Biology and Medicine*, vol. 136, N 4, pp. 325-327, 2003.
- [2] Bondarenko N.S. "Protective Effects of P-G-P in Compound 48/80 Induces Anaphylactoid Reactions", *5th International Symposium on Cell/Tissue Injury and Cytoprotection/Organoprotection*, Yalta, Ukraine, September 17-19, 2008.
- [3] Zolotarev Yu.A., Badmayeva K.E., Bakayeva, et al., "Short peptide fragments with antiulcer activity from a collagen hydrolysis", *Bioorg. Khim.*, vol. 32, N 2, pp. 89-94, 2006.
- [4] Ashmarin I.P., Baglikova K.E., Edeeva S.E. et al., "Oligopeptides: pharmacokinetics", *Bioorg. Khim.*, vol. 34, N 4, pp. 464-470, 2004.
- [5] Cherkasova K.A., Lyapina L.A., Ashmarin I.P., "Comparative study of modulatory effects of Semax and primary proline-containing peptides on hemostatic reactions", *Bull Exp Biol Medical*, vol. 132, N. 1, pp. 625–626, 2001.
- [6] Martinova K.B., Andreeva L.A., Klimova P.A. et al., "Structure functional investigation of the glysin and prolin containing peptides which are neuroprotectors", *Bioorg. Khim.*, vol. 35, N 2, pp. 165-171, 2009.
- [7] Akhmedov N.A., Gadjiyeva Sh.N., Abbasli R.M. "Structural organization of Asp-Pro- Lys-Gln-Asp-Phe-Met-Arg-Phe-NH₂ molecule", *Cur.Top. in Pept.&Prot.Res.*, vol.10, pp. 57-62, 2009.
- [8] Akhmedov N.A., Ismailova L.I., Agayeva L.N., Gocayev N.M. "The spatial structure of the cardio active peptides", *Cur.Top. in Pept.&Prot. Res.*, vol.11, pp. 87-93, 2010.
- [9] Akhmedov, N.A., Ismailova, L.I., Abbasli, R.M., Akhmedov, N.F., and Godjaev, N.M. "Spatial structure of Myelopeptides: I. Conformational analysis of MP-1, MP-2, and MP- 3", *Bioorg. Khim*, vol. 31, pp. 1-7, 2005.
- [10] Popov, E.M. "An Approach to calculations of the problem of structure-functional organization of natural peptides", *Mol. Biol.*, vol. 19, pp. 1107-1138, 1985.
- [11] Maksumov, I.S., Ismailova, L.I. and Godjaev, N.M. "A computer program for calculation of conformations of molecular systems", *J. Struc. Chem.*, vol. 24, pp. 147- 148, 1983.
- [12] IUPAC-IUB, *Quantity, Units and Symbols in Physical Chemistry*, vol. 39, Blackwell Scientific Publications, Oxford, 1988.

ZIPPER: The Holistic Spell Checker

Lina Alhusaini
School of Informatics
The University of Edinburgh
EH8 9AB Edinburgh, United Kingdom
Email: s0460867@sms.ed.ac.uk

Abstract—This article discusses a novel problem with a novel solution. The thesis of this work is to perform document and text editing. We look specifically at providing a document editor with a tool for text editing and that tool is *holistic spell checker*. We propose ZIPPER, a holistic spell checker, that finds out user misspell words *all at once*. Our approach to this approximated problem is probabilistic. We use Markovian tree that exploits the dependencies amongst characters in a word. For computation over Markovian tree, we use a set of probabilistic and information theory metrics. For link quantification, we use information theory metric which is *pointwise mutual information*. Probabilistically, we use *belief propagation using message passing paradigm* for node quantification. To create a suggestion list for each misspelled word, we decompose Markovian tree into clique tree. For computation over clique tree, where each node is a complete word, we use information theory metrics like: *entropy* for computing the value of a complete word, and *mutual information* for computing how much value there is between two words.

I. INTRODUCTION

Spelling is mainly used in document and text editing, like: Microsoft Office. Misspelling occurs by user either due to typing errors or due to lack of knowledge of correct spelling. In small size documents, checking the misspelled words, word-by-word, is simple. Yet, the task of spell checking gets harder in large size documents containing thousands of words. The case of large documents motivated the act of *holistic spell checking*. This solution eases the correction of misspellings for large number of words and it is the invention of this article.

A. The Problem

The problem being solved in this work is an unapproached problem. It is the case of holistic spell checking. By definition [1], the word *holistic* means whole part. This appears in our work by spell checking many words (whole part) at once. Even though we can spell check only one word, the emphasis is on spell checking holistically; that's two words and more. We give the name ZIPPER to this work for its ability to tune to one tick, two, or as many as there are. The act of a zipper simulates spell checking words either one word, two, or as many as there are.

II. LITERATURE SURVEY

The task of spell checking is derived from linguistics. As in the science of languages there are three sections, the main part considering spell checking in language form. Within it, Saussurean Sign, the classical model of linguistics sign, is divided into two parts: shape of a word and its phonics. In

the shape of a word part, that is known as morphology, it recognizes the structure of a language in terms of its words, affixes, parts of speech, etc. The syntax of a language is in regard to the words building sentences. Spell checking is mainly concerned with a language words.

Spelling is a tool for document and text editing. It appears is so many of today's softwares, like: Microsoft Office, latex different platform softwares, dictionaries, Adobe softwares, messengers, yet to name a few.

The author of the work in [2] has clearly defined three types of spell checking research: (a) Nonword error detection research; (b) Isolated-word error correction research; and (c) Context-dependent word correction research. For the first research, it has spanned two decades from the 1970s to the 1980s and used two techniques: n-gram analysis and dictionary lookup. For the second research, interestingly, it started in the 1960s until present. This research does not only detect errors, it also tries to correct the errors. It does so either automatically without user intervention like in text-to-speech recognizers or manually by user. Intensive investigations being devoted for this type of research. For the third research, it started early 1980s with techniques from natural language processing.

III. BACKGROUND

Some background information for the reader.

1. Probability [3]. The field of probability measures the amount of certainty of the likelihood happening of a condition. It is used to solve problems approximately.

(a) Markov Model. This model is a probabilistic model. It exploits the dependencies amongst nodes in a graph to compute marginal probability. It supports tree representation.

(b) Belief Propagation Using Message Passing Paradigm. It is a tool accompanied with Markov Model to compute nodes marginal probability (a.k.a. belief) in a graphical representation using messages.

(c) Clique Tree. It is a tool accompanied with Markov Model. It is a decomposition model of the original tree in the algorithm behavior model. A clique tree nodes are the original values entered in the original tree of the behavior model.

2. Information Theory [4]. We use information theory concepts in the study of individual characters of words. We try to understand the meaning of characters position and

sequencing.

(a) Entropy. Entropy is a measure of uncertainty firstly used in signal processing to quantify the loss of the amount of information to the other site. For example, typing the characters of a word in a dictionary also has an amount of loss of certainty either due to typing errors or due to lack of knowledge of correct spelling.

(b) Mutual Information (MI). MI is the reduction in uncertainty of variable X due to the knowledge of variable Y. Given Figure 1, X and Y are FLOWER and FLOR, consequently. And MI computes the intersection value between them.

(c) Pointwise Mutual Information (PMI). PMI is a measure of uncertainty of points in a distribution. It computes the amount of information of event y given event x. Given Figure 1, PMI computes a value for point o.

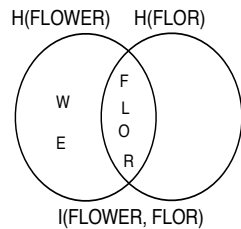


Fig. 1. Venn Diagram of Two Intersecting Words

3. Parallelization [1].

It is the science of dividing large problems into smaller ones and solve them concurrently. There are four types of parallelism: bit-level, task, data, and instruction-level parallelism.

(a) Data Parallelization. It divides data into smaller chunks and distributes them over multiple processing nodes.

(b) PCAM Strategy. It is a model to design parallel algorithm where four elements are essential: Partitioning of computation and data; Communication of tasks concurrently; Agglomeration of tasks for efficiency; and Mapping tasks on different computing processors.

IV. THE TOOLBOX

We present the tools required for the design of holistic spell checking algorithm [5]: (1) Problem Solving Type, (2) Algorithm Design Technique, (3) Data Structures, and (4) Computational Means.

(1) Problem Solving Type. A problem solving type questions the preconditions for solving a problem. There are two types of solving problems. One is exact solving, and another is approximate. If all preconditions of a problem are satisfied, then the problem solving type is exact. If there are one or two preconditions not satisfied, then the problem will be solved approximately. Most approximate problems are NP hard problems in that there is not an exact solution for it. The nature of the problem being solved in this work is approximate. That is, the output of the program provides approximate answers within which an exact answer can sometimes be found and in other times can not be found. By approximation we mean that there are a set of preconditions that are not all satisfied. In this work, the precondition that is not satisfied is a correct spelling of user words. Approximation appears by our usage of probability.

(2) Algorithm Design Technique. “An algorithm design

technique (or “strategy” or “paradigm”) is a general approach to solving problems algorithmically that is applicable to a variety of problems from different areas of computing” [5]. That is to say, to solve problems you need a “design idea” for developing an algorithm for it. Design ideas may includes: (a) Brute Force; (b) Divide-and-Conquer; (c) Decrease-and-Conquer; (d) Transform-and-Conquer; (e) Space and Time Tradeoffs; etc. Every design idea is a *key* to solve a problem by choosing an applicable algorithm design technique to develop a solution algorithm. Knowing your algorithm design technique is important for two reasons: (a) Guidance: you will have a design idea that will help you develop your solution algorithm; and (b) Classification: the developed algorithm will have a class of its own for reference. For this work, we choose “Space and Time Tradeoffs.” This algorithm design technique uses extra space to speedup solution formation. There are a number of types for this technique. They are: (a) input enhancement; (b) pre-structuring; and (c) dynamic programming. From these types we choose pre-structuring. The reasons behind this choice are its ability to combine both (a) pre-processing; along with (b) fast and/or (c) flexible access to the data. In this work, the chosen design technique is materialized through the usage of Hashing.

(3) Data Structures. Data structures are means that hold data for processing. Any problem would require at least one data structure. There are a varieties of data structures like: array, queue, stack, linked list, etc. Data structures are chosen targetly for a particular problem. We choose the following data structure to solve our problem.

Tree (Forest of Rooted Trees): We specifically choose a rooted tree data structure and into the creation of a forest of rooted trees which intuitively model our problem. To justify the intuition of our choice, we say: (a) a tree has a root: the root signals the start of a word, intuitively. It also indicates that words with the same first character live on the same rooted tree. As we are modeling words of different first character, we end-up with a set of disconnected rooted trees. Those disconnected rooted trees are known as a forest of rooted trees; (b) a tree has a branch: every character of a word sets as a vertex building a branch. Similar words lie on the same branch. This branch feature of a tree is intuitive to model the words’ characters and to permit the study of each character in isolation; (c) a tree has a leave: a leave indicates the end of a word, except for those words that are a subset of another word, where they do not reach a leave.

(4) Computational Means. Computational means describe how the computation of a certain algorithm is done. We ascertain the computational capabilities by explaining: (a) computational device; (b) computational programming paradigm; and (c) computational model.

A. Computational Device: As this work application is targeted towards a personal use, we choose single processing architecture computers which are so called RAM; Random Access Machines.

B. Parallel Programming Paradigm: As we deal with large data size, this means our problem is inherently a data

TABLE I
ALGORITHM DESIGN PSEUDOCODE WITH EXPLANATION

Spell Checker Parts	Data Parallization Strategy (PCAM)	English Language and Mathematical Style Pseudocode
Scanning Routines	<p>PARTITION</p> <p>COMMUNICATION</p> <p>AGGLOMERATION</p>	<p>INPUT MODEL: fault-rate</p> <pre> input user words input correctly spelled words merge sort user and correctly spelled words build a forest of trees with user words and correctly spelled words while user words if user word correctly spelled then print fault-rate 1 and remove word from forest else if word is creating a branch for itself then print fault-rate 2 and remove word from forest else if user word is misspelled then print no fault: misspelled user word remains in the forest </pre>
Spell Check Algorithm	<p>MAPPING</p>	<p>BEHAVIOR MODEL: markov model on trees</p> <pre> In parallel over the trees in the forest apply markov model link quantification: $PMI = \log \frac{P(x,y)}{P(X)P(Y)}$ node quantification: $b_i(x_i) = k\Phi_i(x_i)\prod_{j \in N(i)} m_{ji}(x_i)$ $m_{ji}(x_i) = \sum_{x_j} \Phi_j(x_j)\Psi_{ji}(x_j, x_i)\prod_{k \in N(i) \setminus j} m_{ki}(x_i)$ </pre> <p>OUTPUT MODEL: clique tree</p> <pre> build clique tree of extracted user words and closely related correctly spelled words from the forest of trees compute ENTROPY per node in the clique tree $H(X) = -\sum_{x \in X} f(x)\log f(x)$ Compute MUTUAL INFORMATION per a pair of nodes; one is correctly spelled word, and another is user word $I(X;Y) = \sum_{x \in X} \sum_{y \in Y} f(x,y)\log \frac{f(x,y)}{f(x)f(y)}$ decompose clique tree after threshold print suggestion list per user word </pre>

parallel application. For designing the parallel algorithm, we choose PCAM data parallel strategy which stands for: Partition; Communication; Agglomeration; and Mapping. In our case this work is intended to run on a single user computer, which means a single-processor machine. According to Flynn's taxonomy, this type of parallel system is known as Single Instruction Single Data von Neumann architecture (SISD.) The parallel programming language to use is Java. Specifically, we use threads technique to stimulate the nature of a multi-processor parallel system. We choose Java for its cross-platform feature.

C. Markov Model: We choose to apply Markov Model as a probability distribution on the trees for the following reasons: (a) Probabilistic: this work problem of finding that a word is misspelled is inherently probabilistic; (b) Independency of states: that is Markov property that permits the study of a single character independently; (c) Markov Model Tools: tools that are useful to compute marginal probabilities of vertices of the trees, such as: Belief Propagation using Message Passing Paradigm. Other useful tools for computing solutions from the trees, i.e. extracting words after links and nodes quantification, are: Clique Tree Construction and Clique Tree Decomposition.

V. ALGORITHM DESIGN

The design of the algorithm is given in table I. As for any spell checker, there are scanning routines that scan and extract words from file. There is also a spell checker algorithm that finds suggestion word(s) for user misspelled word(s). As in the pseudocode, input model is the first part of the spell

checker. Behavior and output models are the second part of the spell checker. Regarding data parallelization strategy, given in the second column, there are four steps in the PCAM data parallelization strategy: Partition - where data is split into chunks; Communication - data chunks communication; Agglomeration - combining some chunks of data prior processing; and Mapping - decision of what chunk should be processed on what processor. As in the pseudocode, the first three steps of PCAM strategy is accomplished in the *input model*. The fourth step, mapping, is accomplished by the *behavior model*.

For writing the algorithm pseudocode style, we use English language, programming *while* for looping, and *if-then-else* for conditioning accompanied with mathematical style pseudocode. The algorithm three basic models are: (a) input: fault-rate; (b) behavior: markov model on trees; (c) output: clique tree.

For Input Model, we choose Fault-Rate Model to outline the misspelled words in user data. We firstly read user data, and then correctly spelled list of data. We sort both user and correctly spelled data using merge sort for its efficiency with large scale data. From correctly spelled data, we create a forest of rooted trees sorted alphabetically. We also pass user data through the forest of rooted tree. After conditioning user data, words that are correctly spelled or creating a branch for itself are extracted from the forest of trees. Other user data remain in the trees as misspelled user data.

For Behavior Model, we choose Markov Model on Trees. As our problem is probabilistic in its intrinsic, we found

that Markov Model is best fit for it due to its support for a tree graphical representation. In this model, we parallelize execution over the forest of rooted trees. In each tree, we first compute the probabilistic distribution over the tree. We compute independent probability per node. Also, we compute conditional probability per pair of nodes. Finally, we compute joint probability. Once accomplishing the probabilistic distribution, we then compute link quantification using information theory metric and it is pointwise mutual information (see background section 2.c). From there, we compute node quantification, a.k.a belief, using belief propagation using message passing paradigm (see background section 1.b).

For Output Model, we choose Clique Trees. Clique trees are decomposed trees of user words and correctly spelled words. In these trees, we compute entropy (see background section 2.a) per word to know how much value there is in a word. Then, prior to deciding the suggestion list, we compute mutual information (see background section 2.b) per a pair of nodes. Then, using the mutual information value in thresholding, we decompose the clique tree and build the suggestion list per user misspelled word.

VI. ALGORITHM ABSTRACT PROOF

The type of theorem of our problem is of form deductive. We use the forward-backward method [6] that fits all proofing techniques for different theorem types. We use this proofing method abstractly due to the early stage of the project.

The first step in any proof is to recognize: what is given as true (A); the hypothesis, and what needs to be proved true (B); the conclusion. In this work, we use the IF-THEN construct to point-out A which comes after IF and before THEN, and to point-out B which comes after THEN.

This work problem in an IF-THEN construct:

IF a spell checker application program with holistic algorithm, **THEN** ZIPPER is the spell checker with holistic algorithm.

Pointing-out A and B, we get:

A: A Spell Checker Application Program with a Holistic Algorithm

B: ZIPPER is the Spell Checker Application Program with Holistic Algorithm

Applying the forward-backward method, we start by moving backward and use the information contained in A that links with B. The abstraction process constitutes three parts: (a) Abstract Question; (b) Abstract Answer; and (c) Specific Answer. By starting the abstraction process over B, we get:

Abstract Question: how can I show that ZIPPER is a spell checker application program?

Abstract Answer: we need to show that ZIPPER is capable of spell checking a word against a correctly spelled list of words.

Specific Answer: $\text{ZIPPER}(w) = \text{Spell-Checks}(w, CW)$; this means that when ZIPPER reads a word (w), it spell checks it against a correctly spelled list of words (CW).

The abstraction process over B has given us a new statement B1 with the property that if B1 is true then B is also true.

B1: $\text{ZIPPER}(w) = \text{Spell-Checks}(w, CW)$
Now it is time to prove that B1 is true and that's by applying

the abstraction process over B1. If we can show that ZIPPER is a spell checker, then we can prove that ZIPPER uses a holistic algorithm, which is capable of spell checking many words at once.

Abstract Question: how can I show that ZIPPER uses a holistic algorithm in its spell checking application program?

Abstract Answer: we need to show that ZIPPER is capable of spell checking more than one word at a time against a correctly spelled list of words.

Specific Answer:

$\text{ZIPPER}(w1, w2) = \text{In-Parallel}(\text{Spell-Checks}(w1, CW), \text{Spell-Checks}(w2, CW))$; this means that when ZIPPER reads more than one word - word 1 and word 2 (w1,w2) -, it spell checks them in parallel against a correctly spelled list of words (CW).

The abstraction process over B1 has given us a new statement B2 with the property that if B2 is true then B1 is also true.

B2: $\text{ZIPPER}(w1, w2) = \text{In-Parallel}(\text{Spell-Checks}(w1, CW), \text{Spell-Checks}(w2, CW))$

As we have let the information in B guides us in formulating the aforementioned abstraction questions, there seems no more information to utilize. Thus, we now need to move forward.

As we are trying to prove that A implies B, we are allowed to assume A is true. It is time to do so through the forward process.

The forward process starts over statement A which we assume is true and later derive statement A1 which is also assumed to be true.

A: A Spell Checker Application Program with Holistic Algorithm

As statement A contains two main information and they are: (a) spell checker application program, and (b) holistic algorithm. Let us now write statement A1 which utilizes information (a), and statement A2 which utilizes information (b).

A1: $\text{Spell-Checker}(w) = \text{Spell-Checks}(w, CW)$
A1 states that when a spell checker application program reads one word, it spell checks it against a correctly spelled list of words.

Deriving statement A2:

A2: $\text{Spell-Checker}(w1, w2) = \text{In-Parallel}(\text{Spell-Checks}(w1, CW), \text{Spell-Checks}(w2, CW))$
A2 states that if more than one word is read through the spell checker application program, then the processing will be done in parallel.

By substitution, we now derive statements A3 and A4. To derive statement A3, we substitute B1 into A1 to get:

A3: $\text{Spell-Checker}(w) = \text{ZIPPER}(w)$
Statement A3 states that ZIPPER is a spell checker application program and that is the first information we need to prove true.

Next, to derive A4, we substitute B2 into A2 to get:

A4: $\text{Spell-Checker}(w1, w2) = \text{ZIPPER}(w1, w2)$
Statement A4 states that ZIPPER has a holistic algorithm that spell checks more than one word in parallel and that is the second information we need to prove true.

Statements A3 and A4 shows that ZIPPER is true and its building is feasible. **Q.E.D.**

VII. ALGORITHM ANALYSIS

Asymptotic big-oh notation is used for the algorithm time and space complexity analysis. We use Random Access Machine (RAM) model for computing the complexity of time and space for the Input and Output model of the algorithm. We use Parallel Random Access Machine (PRAM) model for the parallel algorithm of the Behavior model for computing the time and space complexity analysis.

1) *Input Model: Fault Rate: Extraction.* User words and correctly spelled words are extracted from files for processing. The time and space consumption is proportional to the words read. For example, if n is the total number of user and correctly spelled words, the time and space complexity is $O(n)$; polynomial.

Sorting. Merge sort is used to sort user words and correctly spelled words prior to insertion in the tree. Merge sort is used due to its efficiency with large data sets than other sorting algorithms. Its time complexity at best, average and worst case is logarithmic; that's $O(n \log n)$. The space complexity of the recursive merge sort algorithm is $O(n + \log n)$, where n is for the working array and $\log(n)$ is for the stack space which is required for any recursive algorithm.

Tree Creation. The time complexity is $O(|V| + |E|)$ and space complexity is $O(|V|)$ where $|V|$ stands for the number of vertices in the tree which is equal to the number of characters in words. $|E|$ stands for the number of edges between characters in words.

While-loop. The time and space complexity for the while loop with three conditions is proportional to the number of user words (n); $O(n)$.

2) *Behavior Model: Markov Model:* PRAM model is used in the complexity analysis of this behavior model. Threads in Java language is used to stimulate the behavior of multiple processes machine. So, the analysis below is per process; i.e. thread.

Probability Distribution. Computation of individual characters probability, conditional probability, and joint probability is distributed over the forest of rooted trees. Each probability computation is proportional in time and space to the number of characters read.

Link Quantification. To quantify the links using the point-wise mutual information (PMI), we need to visit each link once. Thus, time and space complexity is $O(|E|)$, where $|E|$ stands for the number of edges in the tree.

Node Quantification. As we use Belief Propagation using Message Passing Paradigm to quantify nodes, we note that it behaves like depth first search (DFS) algorithm with repetition equals to the number of vertices. So, time complexity at worst case is $O((|V|+|E|)*|V|)$, where $|V|$ stands for the number of vertices of a tree and $|E|$ for the number of edges in the tree. Space complexity is $O(|V|*|V|)$.

3) *Output Model: Clique Tree: Clique Tree Construction.* The time complexity is $O(|V| + |E|)$ and space complexity is

$O(|V|)$ where $|V|$ stands for the number of vertices in the tree which is equals to the number of words. $|E|$ stands for the number of edges between closely related words.

Clique Tree Decomposition. The time to decompose the clique tree is constant and is performed after threshold. No extra space is consumed.

Suggestion List Production. Time is polynomial to the number of user words $O(w)$, in addition to the correctly spelled suggested words per misspelled user word $O(sw)$. Space complexity is proportional to the total number of words; both user and suggested words.

VIII. RELATED WORK

There exists a number of spell checkers [1], like: GNU Aspell, Ispell, MySpell, and Pspell. And a number of language-specific spell checkers, like: Hunspell for Hungarian, Voikko for Finnish, Zemberek for Turkish, and Virastyar for Persian. Each is designed with a specific orientation, yet all are designed to spell check a word at a time. While ZIPPER spell checks many words at once.

IX. CONCLUSIONS

We presented in this paper a novel problem with a novel solution. It is the problem of holistic spell checking which arises in the context of document and text editing. In this work, we discuss holistic spell checking algorithm from the following viewpoints: toolbox, design, abstract proof, and analysis. From the discussion in this article, we conclude that the concept of holistic spell checking is feasible.

ACKNOWLEDGMENT

To the knowing of knowledges.

REFERENCES

- [1] <http://wikipedia.org>
- [2] K. Kukich, *Techniques for Automatically Correcting Words in Text*, ACM Computing Surveys, 24(4), 1992.
- [3] J. Pearl, *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*, Morgan Kaufmann, 1988.
- [4] H. Scutze and C. Manning, *Foundations of Statistical Natural Language Processing*, MIT Press, 1999.
- [5] A. Leviton, *Introductions to the Design and Analysis of Algorithms*, Pearson, 2003.
- [6] D. Solow, *How to Read and Do Proofs: An Introduction to Mathematical Thought Process*, John Wiley and Sons, 1982.

Analysis of the Relation between Turkish Twitter Messages and Stock Market Index

Mehmet Ulvi Şimşek
Computer Engineering Department
Faculty of Engineering, Gazi University
Ankara, TURKEY
mehmetulvi@gmail.com

Suat Özdemir
Computer Engineering Department
Faculty of Engineering, Gazi University
Ankara, TURKEY
suatozdemir@gazi.edu.tr

Abstract— The increasing popularity of social networks also increased the amount of data collected in these networks. Online social applications, which are also defined as social media, provide ways of accessing large amounts of data about their users over the Internet. Not surprisingly, researchers started to focus on the data extraction process in these networks. In this context, data mining applications and data analysis allow researchers to extract some useful information about the masses and people. In this study, we examine a Turkish twitter data set using data mining techniques. Initially, Turkish tweet dataset is collected and emotional words are determined. An analysis is carried out to see if there is a relation between Turkish tweets and the Turkish stock market index. To the best of our knowledge, this is the first study performed on Turkish tweets and stock market index.

Keywords- Data mining, social networks, twitter, stock market

I. INTRODUCTION

Applications developed on social networks give people not only communication and information share but also entertainment and the opportunity to spend good time. Applications, which are defined as social media, provide ways of accessing large amounts of data about people and the masses over the Internet. These data sources on the Internet can be listed as blogs, information sharing sites, online gaming sites, news groups and chat rooms [1,4]. Today, facebook [10] and twitter [11] are the most used social networks. Blogs have an important role in the data source of social networks as well [15]. Blog sites whose users can write comment on specific issues has become much more effective via social tagging systems developed [4]. Users can enter these sites at any time and communicate with each other or explain what you think about at any time. This behavior which is referred as collective behavior is followed by sprawling structure [14]. For this reason, a very large of data accumulated on social networks.

Useful information can be extracted via data mining methods by examining the data left by users of social networks. Social networks or particular groups can be classified based on emotional classes via examination of short articles on sites such as twitter users. Determining the word and language properties is an important step in social network analysis. As all languages have different properties, analysis

techniques should be different as well. For example, in Chinese language, you can analyze the characters instead of words [13].

In this paper, we use the corpus of emotional keywords in Turkish language and choose the most common Turkish words to investigate how the public emotions change. Using Turkish twitter dataset, we are able to examine the how happy the users are. We extract the most common words from emotional keyword dataset. We analyze the correlation between stock market data and twitter data whose messages containing stock market words.

The rest of the paper is organized as follows. In Section 2, the related work in this field is presented. Section 3 explains the analysis of Turkish tweet messages where Section 4 and 5 present the calculation of average happiness and unhappiness words. In Section 5, the relation between Turkish tweets and the stock market index is shown. Finally, the concluding remarks are given in Section 6.

II. RELATED WORK

Most opinion mining applications are related to electronic commerce and trade [7]. Few studies include the social and geopolitical studies. Opinion mining can be used many fields. For example, researchers use opinion mining to evaluate performance of companies in stock market forums. Another example, people's perception about political figures can be extracted from social forum and sites [7].

Behavioral economics says that emotions affect individual decisions and behaviors. The relationship between social networks and economic situation has been the subject of research [8]. Positive and negative ideas on the change of market shares are examined via the examination of the financial forums [7]. Lately, Twitter platform have been used for this kind of research as twitter users can easily write short text about various subject such as economics.

The 'happiness' level of individual words are used in research to investigate public happiness, songs, blogs etc [6,8]. Hedonometer algorithm [6] is used for evaluate to short text messages and it gives to researchers to obtain weighted

happiness level for the tweet text [6]. The average happiness values for words which is evaluated by humans used to obtain happiness level for the text or twitter text [6,9].

Language characteristic is important for the author to analyze the changes in public opinion. Most commonly used words in the Internet must be preferred. Many researchers use the emotion corpus {Anger, Sadness, Love, Fear, Disgust, Shame, Joy, Surprise}[5]. Recently, researchers give great attention to opinion mining by using emotion corpus which helps classification of the public opinion [3]. Similarly, tweets are classified by using the author information and features within the tweets [2].

III. THE ANALYSIS OF TURKISH TWITTER MESSAGES

In this study, first, the words used for analysis of Turkish twitter messages selected. Turkish tweet messages are collected from twitter via Twitter API [12].

A. Description of Dataset

We have collected tweets over a 45-day period between 16.12.2011 00:00 am and 31.01.2012 24:00 pm. Our data set includes approximately 1.9 million tweets. The collected tweets are in Turkish.

B. Selection of Happiness and Unhappiness Words

Emotion corpus has 8 class {Anger, Sadness, Love, Fear, Disgust, Shame, Joy, Surprise} [3]. We have select 113 words in Turkish from emotion corpus to form the following two classes: {Happiness, Unhappiness}. We use the most common Turkish words which are thought to represent happiness and unhappiness.

C. Analysis of Happiness and Unhappiness

Initially, we examine the tweet dataset for predetermined words indicating happiness and unhappiness. If such a word found in a tweet, we set the tweet vector 1, otherwise 0. We compute the percentage of happiness tweets by using happiness word frequency. Similarly the percentage of tweets indicating unhappiness is computed. The following formulas show the computation and Figure 1 and 2 show the daily percentage of happiness and unhappiness of the Turkish tweets.

$$\frac{\sum_{i=0}^k \text{happy words}}{\sum_{i=0}^n \text{tweet_count}}, \quad k = \text{happy words count for each tweet}$$

$$\frac{\sum_{i=0}^l \text{un happiness words}}{\sum_{i=0}^n \text{tweet_count}}, \quad l = \text{unhappy words count for each tweet}$$

$$\frac{\sum_{i=0}^m \text{other}}{\sum_{i=0}^n \text{tweet_count}}, \quad n = \text{tweet count}$$

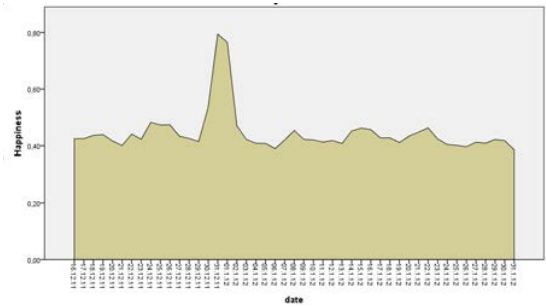


Figure 1. Change in happiness percentage in time domain graphic

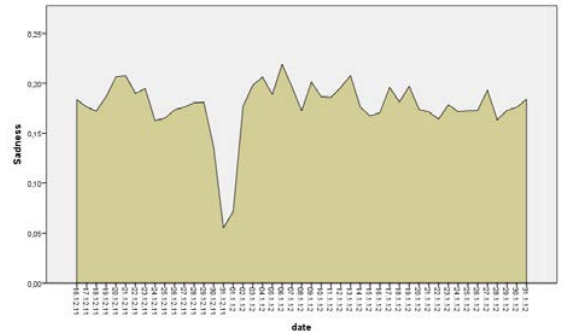


Figure 2. Change in unhappiness percentage in time domain

It is observed from Figure 1 and 2 that between 31.12.2011 and 01.01.2012 the percentage rate of happiness is reached the maximum observed value. This is expected as Twitter users use more word expressing happiness during the New Year 's Eve.

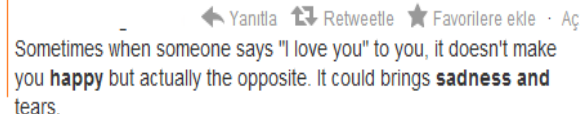


Figure 3. Happy and unhappy words are used together with an example of tweets

However, simple frequency analysis is inadequate to evaluate the tweets which contain both happiness and unhappiness words. Happiness and unhappiness words may be used together as seen in the Figure 3. For this kind of tweets, the average happiness values are used as explained in the next section.

IV. CALCULATION OF AVERAGE HAPPINESS

Average happiness values are used to evaluate public happiness with human evaluations [6]. These words represent a spectrum that varies from sad to happy. The human evaluated average happiness values for the 113 Turkish words of emotion corpus [3] are selected from Mechanical Turk [6]. We use these average values of words to evaluate short text which contain both happiness and unhappiness words. These rates of the words are determined by Mechanical Turk by

asking people their feelings about the words on a nine point integer scale [6]. Some examples of average values are given below. For example, the word “happy” (in Turkish mutlu) receives 8.30 point whereas the word “depression” (in Turkish depresyon) receives 1.98 point.

Average_happiness(“mutlu”)= 8.30
 Average_happiness(“depresyon”)= 1.98

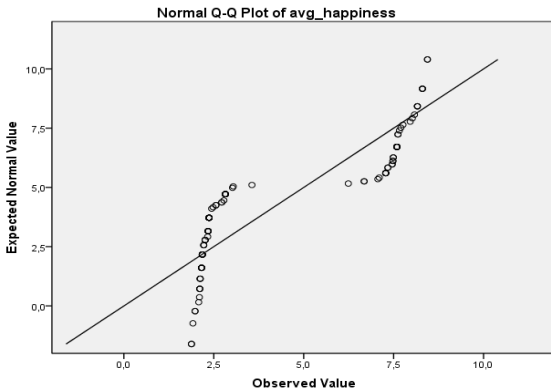


Figure 4. Distribution of words according to average happiness value

Q-Q plot is a graphical method for comparing two probability distributions by plotting their quantiles against each other. Average happiness and unhappiness values of the analyzed words are shown in the Figure 4. The Figure 4 shows each word in a circle. Unhappiness words indicating values lower than 5 are gathered in the upper right corner of the Figure 4, happiness words indicating values greater than 5 are gathered in the upper right corner of the Figure 4.

We use the formula proposed in [6] to evaluate both happiness words and unhappiness words in the same tweet. This effective method gives us to evaluate the each message’s happiness level. This formula is given below.

$$Avg(T) = \frac{\sum_{i=0}^n Avg_{t_{happiness}}(word_i) * f_i}{\sum_{i=0}^n f_i}$$

where f_i = frequency of i ‘th word, $Avg_{happiness}(word_i)$ = Happiness value of word and $Avg(T)$ = average happiness of given tweet.

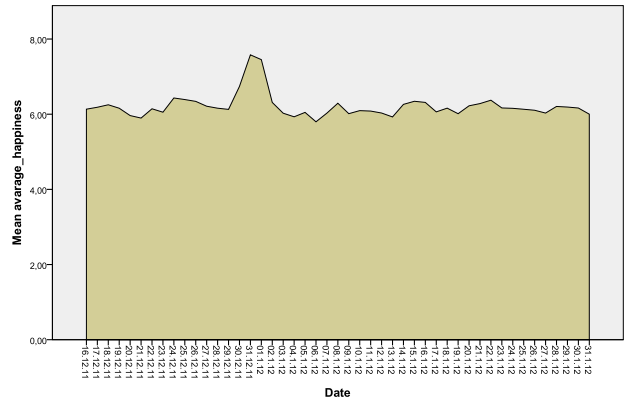


Figure 5. Change in average happiness for tweets in time domain

Changes in the average happiness values of tweets from 47 days are given in Figure 5. The average value of happiness is greater than 5 for each day because the uses of happiness words are greater than unhappiness words. The maximum rate of average happiness is around New Year’s Eve.

V. COMPARISON OF STOCK MARKET DATA AND TWITTER MESSAGES

Economic indicators may affect happiness or unhappiness in social networks. People in the social network can easily react in short time about issues such as economic changes. In twitter, users comment any issue in short time. These comments give us to information and public reaction about the issue. With examining the comments, public reaction can be evaluated. Users in twitter and similar social networks might be affected by the change in the stock market has drawn the attention of researchers.

To evaluate the relation between stock market change and public response, specific tweets are selected. The chosen tweets, which contain the words related to stock market in Turkish, are evaluated with using average happiness value of each tweet. Closing market prices of IMKB100 data are used to evaluate how public response about market prices. When the stock market decrease, the day is determined “unhappiness” and this day tweets happiness rate are expected to below 5. When the stock market increase, the day is determined “happy” and this day tweets happiness rate are expected to above 5. Similarly, selected tweets from database for each day evaluated and each day is determined as “happy” or “sad”. As a result of this examination, it is shown that stock market and tweet data relation is about approximately %45. Average happiness rate are given below in Figure 6.

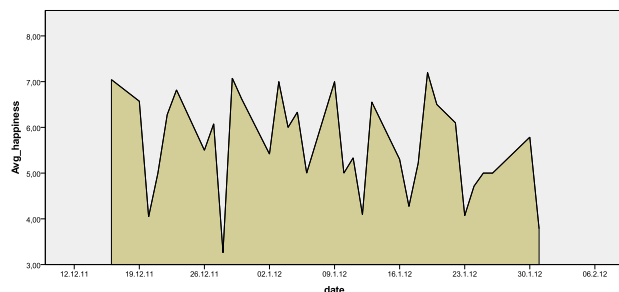


Figure 6. Change in average happiness for tweets in time domain

VI. CONCLUSION

In this study, Turkish emotional words are determined to be used in social network analysis and their happiness and unhappiness values are computed using frequency analysis and average happiness value analysis. Initially, we observe the time domain changes in Turkish twitter messages. It is clearly seen that around the New Year days happiness rate is increased in both the frequency analysis and average happiness value analysis. The words “happy” and “trouble” in Turkish are the most commonly used words in emotional word database. Then, we analyze the relationship between twitter messages and the Turkish stock market index. When the twitter messages which contain stock market related words analyzed, it is seen that average emotional value of the tweets changes from happy to unhappiness. Stock market and tweet data relation is found to be approximately %45.

We hope that this emotional database in Turkish and analysis of happiness will be used by other researchers. As our future work, by expanding the emotional word list, we plan to examine the public response about the specific issues such as politics and sports events.

ACKNOWLEDGMENT

This work is supported in part by the Gazi University Scientific Research Project Funds No. 06/2011-41.

REFERENCES

- [1] P. Domingos, “Mining Social Networks for Viral Marketing”, *IEEE Intelligent Systems*, vol 20(1), pp. 80-82 (2005).
- [2] M. Demirbaş, B. Sriram, D. Fuhry, E. Demir and H. Ferhatosmanoğlu, “Short Text Classification in Twitter to Improve Informative Filtering”, *Proceeding of the 33rd international ACM SIGIR conference on research and development in information retrieval*, Geneva Switzerland, pp. 841-842, 2010.
- [3] C. G. Akcakora, M. Bayir, M. Demirbaş and H. Ferhatosmanoğlu, “Identifying Breakpoints in Public Opinion”, *SOMA*, Washington, pp. 62-66, 2010.
- [4] M. U. Şimşek, S. Ozdemir and H. Karacan, “Data Mining in Social Networks”, *Bilişim 2011*, 26-28 October, Ankara Turkey, 2011.
- [5] W. G. Parrott, editor.” *Emotions in social psychology: essential readings*”, Psychology Press, 2001
- [6] P. S. Dodds, K. D.Harris, I. M. Kloumann, C. A. Bliss and C. M. Danforth, “Temporal patterns of happiness and information in a global social network: Hedonometrics and Twitter”, *Plos One*, vol 6 (12), 2011
- [7] H. Chen and D. Zimbra, “AI and Opinion Mining”, *IEEE Intelligent Systems*, vol 3 (25), pp. 74-76, 2010.
- [8] H. Mao, J. Bollen and X. J. Zeng, “ Twitter mood predicts the stock market ”, *Journal of Computational Science*, vol 2 (1), pp. 1-8, March 2011
- [9] P. S. Dodds and C. M. Danforth, “Measuring the Happiness of Large-Scale Written Expression: Songs, Blogs, and Presidents”, *Journal of Happiness Study*, vol 11 (4), pp. 441-456, 2010.
- [10] Web: www.facebook.com, 2012
- [11] Web: <http://twitter.com/>, 2012
- [12] Web: <https://dev.twitter.com/>, 2012
- [13] L. Ku, Y. Liang, and H. Chen, “Opinion extraction, summarization and tracking in news and blog corpora”, In *Proceedings of AAAI-2006 Spring Symposium on Computational Approaches to Analyzing Weblogs*, Palo Alto USA, pp. 100-107, 2006.
- [14] H. Liu and L. Tang, “Toward Collective Behavior Prediction via Social Dimension Extraction”, *IEEE Intelligent Systems*, vol 25(4), pp. 19-25, July/August 2010.
- [15] H. Liu, J.J. Salernox, N. Agarwal, P. S. Yu and S. Subramanya, “Connecting Sparsely Distributed Similar Bloggers”, *ICDM '09. Ninth IEEE International Conference*, Miami USA, 2009.

Computer Determination of Preferred Conformations of Human Hemokinin-1

U.T.Agaeva¹, G.A. Agaeva², N.M.Godjaev^{1,2}

1-Institute for Physical Problems, Baku State University ,

AZ-1148, Baku, Z.Khalilov Str.23, Azerbaijan

2-Qafqaz University, AZ-0101,Baku -Sumqait Road,16 km, Azerbaijan

gulshen@mail.ru

Abstract- The conformational properties of biologically active hemokinin-1 peptide molecule have been investigated by computer modeling methods. It is showed that this molecule can exist in several stable conformational states. The energy and geometrical parameters for each of low-energy conformations are obtained. The conformationally rigid and labile segments of this molecule were revealed.

Keywords: hemokinin-1, structure, function, conformation, peptide

I. INTRODUCTION

The scientific interest to the structure and function of the small biologically active peptides is very has increased in recent years, mainly because of the potential pharmacological use of these molecules. Human hemokinin-1 (hHK-1) is an undecapeptide with the sequence Thr-Gly-Lys-Ala-Ser-Gln-Phe-Phe-Gly-Leu-Met-NH₂. This molecule is mammalian tachykinin peptide preferently expressed in peripheral tissues. Human hemokinin-1 (h HK-1) and its carboxy-terminal fragment h HK-1(4-11) are encoded by the recently identified TAC4 gene in human. It is involved in multiple physiological functions such as inflammation, hematopoietic cells development and vasodilatation via the interaction with tachykinin receptor neurokinin-1 (NK1) [1-4].

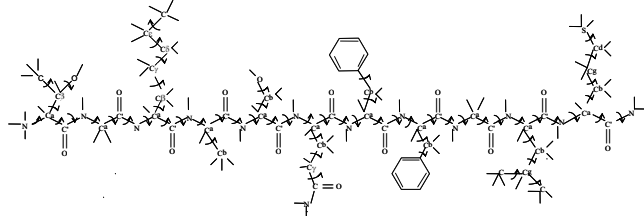


Fig.1.The atomic model of the human hemokinin-1 molecule.

Determination of the mechanism of peptide biological effect and molecular basis of their action is a one of the most important problem of the molecular biophysics. The diversity of biological functions of peptide molecule is undoubtedly connected to its conformational properties. Because the small peptide molecules are very flexible in the aqueous solution. In order to elucidate the mechanism of action of the peptide the

investigation of the native three dimensional structure is necessary, that first of all requires the information about of the full set of low energy and consequently the potentially and physiologically active conformations of this molecule. The major aim of the present article is the investigation of the three-dimensional structure and conformational flexibility for human hemokinin-1, with the purpose of getting insight into basic structural requirement that determine ligand-receptor interaction. The conformational properties of human hemokinin-1 have been investigated by molecular mechanic method, which allow to determine a whole sets of energetically preferred conformers of peptide molecule.

II. STRATEGY AND METHOD

Molecular mechanics (MM) study of human hemokinin-1 conformation involves multistaged extensive computations of even-increasing fragments, with a set stable forms of each preceding step used as a starting set in the next step.

Thr1-Gly2-Lys3-Ala4-Se5-Gln6-Phe7-Phe8-Gly9-Leu10-Met11-NH2

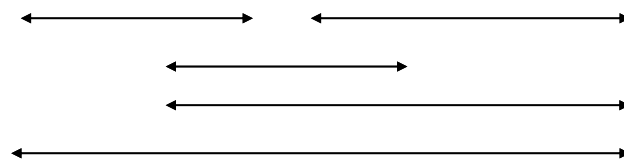


Fig.2..The calculation scheme of the human hemokinin-1 molecule.

Only those conformations are retained whose energies are smaller than some cut-off values. This cut-off value is usually taken as 5 kcal/mol above the lowest energy. The sequential method was used, combining all low-energy conformations of constitutive residues [5]. The conformational potential energy of a molecule is given as the sum of the independent contributions of nonbonded, electrostatic, torsional interactions and hydrogen bonds energies. The first term was described by the Lennard-Jones 6-12 potential with the parameters proposed by Scott and Scheraga. The electrostatic energy was calculated in a monopole approximation corresponding to Coulomb's law

with partial charges of atoms as suggested by Scott and Scheraga. An effective dielectric constant value $\epsilon = 1$ for vacuum, $\epsilon = 4$ for membrane environment and $\epsilon = 80$ for water surrounding is typically used for calculations with peptides and proteins, which create the effects of various solutions on the conformations of peptides by MM method [6]. The torsional energy was calculated using the value of internal rotation barriers given by Momany et al [7]. The hydrogen bond energy is calculated based on Morse potential Bonding lengths and angles are those given by Corey and Pauling and are kept invariable; the ω angle of the peptide bond was fixed at 180° . The dihedral rotation angles were counted according to the IUPAC-IUB [8]. The conformational energy was minimized using program proposed by Godjaev et al. [9].

III. RESULTS AND DISCUSSION

Conformational study of the undecapeptide human hemokinin-1 was carried basing on the fragmental analysis. In each of fragments were used results of preceding small segment. In turn, stages are divided on consecutively decided structured problems. The atomic model of human hemokinin-1 molecule and its variable dihedral angles are represented on the figure 1. The first stage of calculation included consideration of conformational possibilities of pentapeptide fragments according to the calculation scheme, showed in figure 2. But the initial variants of the small fragments were formed on the base of low-energy conformations of the corresponding mono-peptides. Molecular mechanics study of human hemokinin-1 has shown that its spatial structure may be described by two families of low-energy conformations with identical structure of the C-terminal octapeptide. The C-terminal part (residues 4-11) of the human hemokinin-1 can adopt a partially helical structures, but the N-terminal part is different in each family. Only two low-energy conformations of the human hemokinin-1 are fall in the 0-3 kcal/mol energy interval. It is shown that two preferred conformations have similar C-terminal backbone form and values of the relative energy. The global conformation have β -turn structure at the Lys³-Ala⁴-Ser⁵-Gln⁶ segment also. These β -turns are confirmed by distance between C $^\alpha$ atoms of the *i* and *i*+3 residues ($< 7 \text{ \AA}$). The lowest energy structures of human hemokinin-1 exhibit the most favourable dispersion contacts and therefore may be expected to become the most preferred in a strongly polar medium. Calculation show that the electrostatic interactions play a significant role in all optimal spatial structures of molecule. Theoretical conformational analysis of human hemokinin-1 have been indicated few families of low-energy conformations with similar C-terminal octapeptides. This analysis has shown that human hemokinin-1 can form one global, i.e. the lowest-energy structure, which is consist one β -turn on N-terminal part and α -helical segment on the C-terminal part, formed follow hydrogen bonds: NH(Phe⁸)...CO(Ala⁴), NH(Gly⁹)...CO(Ser⁵), NH(Leu¹⁰)...CO(Ala⁶) and NH(Met¹¹)...CO(Phe⁷). But next low-energy conformation with relative energy 1,4 kcal/mol is

formed fully α -helical structure, characterized by following hydrogen bonds: NH(Ser⁵)...CO(Thr¹), NH(Gln⁶)...CO(Gly²), NH(Phe⁷)...CO(Lys³), NH(Phe⁸)...CO(Ala⁴), NH(Gly⁹)...CO(Ser⁵), NH(Leu¹⁰)...CO(Gln⁶) and NH₂(Met¹¹)...CO(Phe⁷). Figure 3 shows the two lowest-energy structures of the human hemokinin-1 as a result of the molecular mechanics of aqueous environment.

IV. CONCLUSION

Thus, on the basis of conformational studies of human hemokinin-1 molecule it has been suggested that the biologically active conformation of this peptide is turned structure in aqueous solution and α -helical structure at its receptor. The obtained data allow one conclude that, in all structures, where is formed α -helical structure at the C-terminal segment residues have many local minimum. The investigation results therefore indicate that a concrete type of the human hemokinin-1 structure will essentially depend on the conditions under which the given molecule functions.

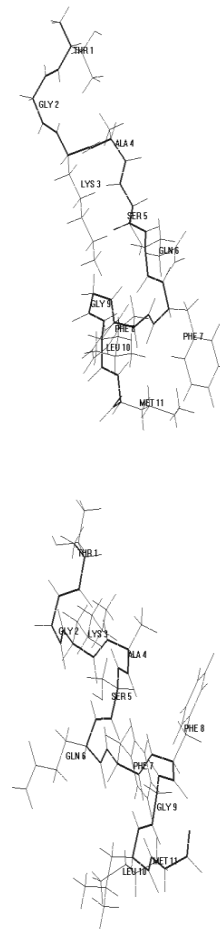


Fig.3. Preferred spatial structures of the human hemokinin-1 molecule.

This work have been fulfilled in the frame of collaboration treaty of the Qafqaz and Baku State Universities.

REFERENCES

- [1] Zhang Y, Lu L, Furlonger C, Wu GE and Paige CJ (2000) *Nat Immunol* 1:392-397.
- [2] Zhang Y and Paige CJ (2003) *Blood* 102:2165-2172.
- [3] Kurtz MM, Wang R, Clements MK, Cascieri MA, Austin CP, Cunningham BR, Chicchi GG and Liu Q (2002) *Gene* 296:205-212.
- [4] Bellucci F, Carini F, Catalani C, Cucchi P, Lecci A, Meini S, Patacchini R, Quartara L, Ricci R, Tramontana M, Giuliani S and Maggi CA (2002) *Br J Pharmacol* 135:266-274.
- [5] Klassert TE, Pinto F, Hernández M, Canden ML, Hernández MC, Abreu J, Almeida TA (2008) *J.Neuroimmunol* 196:27-34.
- [6]] G.A.Agaeva, N.N.Kerimli, N.M.Godjaev, *Biofizika*, vol. 50, 2005, pp. 203-214.
- [7]] G.A.Agaeva, N.N.Kerimli, N.M.Godjaev, *Biofizika*, vol. 50, 2005, pp. 404-412.
- [8] IUPAC-IUB Quantity, Units and Symbols in Physical Chemistry *vol. 39*, Blackwell Scientific Publications, Oxford 1988.
- [9] I.S.Maksumov, L.I.Ismailova, N.M.Godjaev, "The program for semiempirical calculation of conformations of the molecular complexes," *J. Struct. Khim.*, vol. 24, 1983, pp.147-148.

Interactive Teaching Methods of Mathematical Physics by the methods of Computer Visualization

S.T.Huseynov¹, N.V.Ibadov², A.A.Aslanov³
Ganja State University, Ganja city, Azerbaijan Republic
huseynovsahib@gmail, nvibadov@gmail, aaslanov@box.az

Abstract: Interactive learning of mathematical physics at the university education essentially depend on the methods of computer visualization. Currently, there is the big number of software in computer mathematics, and all of them have different quality degree of computer visualization. Therefore, in this direction it is very important to correctly select and properly apply the most suitable software. In the article there are presented different important practical issues of computer visualization and the use of computer mathematics to achieve maximum efficiency in the interactive teaching of the basic divisions of mathematical physics.

I. INTRODUCTION

In virtually all countries, over the last decade, the use of advanced methods of information technology in the teaching of mathematical sciences in university education expands and deepens. This is facilitated by two major technical factors those observed in the development of information and communication technologies (ICT): hardware qualitative changes in modern computers and their peripherals and simultaneously the rapid development of computer mathematics software with their interactive visual properties. It is well known that these factors are very important components in the application of computer technology for interactive learning of mathematical sciences.

In this paper we consider the most suitable software in computer mathematics according to their computer visualization properties to achieve of maximum efficiency in interactive teaching of courses of mathematical physics. The importance of the mathematical physics is tremendous, because this branch of mathematics is most complicated, and in the same time this discipline is very important for other specialties (purely nonmathematical): for students majoring in natural sciences, such as physics, chemistry, biology and Earth sciences.

If in the early of 90-ies the number of software to provide high-quality visualization of scientific data with the mathematical content were no more than dozen, then at the present time, this number already exceeds a hundred. Therefore, at present time it is very important area for professionals, who use software for interactive education of mathematical physics: it is the correct choice among the most suitable software and definition of appropriate use methods of most effective software in learning and teaching. Given our

experience and opinions of various authors [1] and [2], we concluded that in the forefront we have to put forward the practical experience of educator in the interactive teaching with the software. But on the other hand, it is important that he could properly select the corresponding software according to the studied part of mathematical physics. Therefore, in the article, we will adhere to the following: to conduct a critical analysis of the most important software and their internal visual features in accordance with the sections and the methods of mathematical physics. Given the vastness of the list of mathematical programs with interactive visual parameters [3], [4] and [5], we focused on the most popular and affordable software to educators and students to achieve maximum efficiency.

II. VISUALIZATION OF MATHEMATICAL PHYSICS PROBLEMS, SOLVED BY USING THE ANALYTICAL METHODS

Traditional branches of the mathematical physics starts with applying analytical methods of solution: the theory of infinite series, Fourier's separation of variables, the method of characteristics, functions of complex potential, Green's functions, Bessel's functions, integral transforms, theory of operational calculus, methods of perturbation, etc... The importance of these methods, both for practical applications, and for the learning process is undeniable. But what type of the software with its visual characteristics, of course, taking into account primarily of the rich arsenal of analytical packages, is most suitable to the learning process? We believe that the first priority should be given to systems of computer algebra. If you have available the earlier version of Maple, at least 8-th version of Maple, then it is quite enough for interactive visualization to achieve best results in learning process in the mathematical physics, or in another words: in the theory of partial differential equations (PDEs). For analytical calculations in Maple you'll have nice visualization properties, and consequently the best conditions to high printing quality and you can easily transfer your publications to the different text editors. But still, the most appropriate and very important program for interactive learning among the computer algebra systems is the Mathematica software. Analytical calculations in this system are as good as in the Maple, especially for solutions of PDEs. For example look at the (1): it is the linear PDE.

$$\frac{\partial}{\partial x} u(x, y) + x^2 \frac{\partial}{\partial y} u(x, y) = e^x, \quad (1)$$

where $u(x, y)$ - unknown function, x and y are the independent variables.

The symbolical solution of (1) we can find via Mathematica computations in the following Mathematica outputs notation (3).

$$\{ \{ u \rightarrow \text{Function}[\{ x, y \}, e^x + C[1][1/3(-x^3 + 3y)] \}] \}, \quad (2)$$

where $C[1]$ - is the arbitrary function in Mathematica outputs notation.

If to choose the arbitrary function $C[1]$ with specific value, then using Mathematica 3D plotting commands we can immediately to get output of solution in the specific area, for example in range $x(-2, \dots, 2)$ and $y(-2, \dots, 2)$. and . It will be arise as the kind of visualization of 3D graphics of the solution's surface, as it shown below, in the fig. 1.

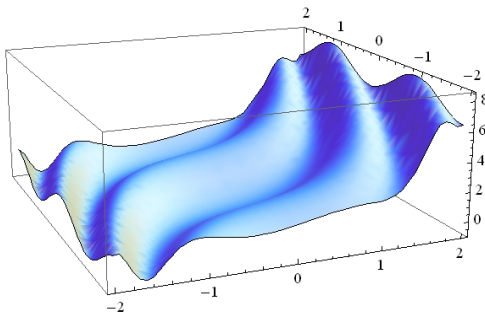


Fig.1. 3D graphical visualization of the symbolic PDE solution surface for a particular choice of the arbitrary function.

III. VISUALIZATION OF NUMERICALLY SIMULATED PROBLEMS OF MATHEMATICAL PHYSICS VIA COMPUTER MATHEMATICS

In the field of computer mathematics for interactive teaching of mathematical physics (or PDEs) the Matlab and Mathematica as computer mathematics software with their high-quality visual settings and interactive features, are the most appropriate tools for the numerical solution of equations of mathematical physics. In the Matlab system for the numerical solution of PDEs there are two powerful tools: Partial Differential Equation Solver (pdepe) and Partial Differential Equation Toolbox. The second of these packages, is more powerful tool that we'll discuss in Chapter IV, as this approach is based on the much more universal numerical method - the finite element method (FEM), which is more closely associated with CAD technology. As for weaker package of PDE Solver, then a standard procedure with the title pdepe in this package - it is a tool for the numerical solution of initial-boundary value problems for systems of parabolic and elliptic PDEs in one space variable and time.

As can be seen from the definition of this numerical of tool, it has a limitations: you can not solve the hyperbolic equations of mathematical physics, as well as 2-D and 3-dimensional problem. And yet, this package is quite possible to use as interactive teaching tool, especially when you consider that in the numerical solution you can control the accuracy of the calculations and the integration step.

As noted above, although now is exist the big number of mathematical software with high quality of visualization parameters of the results of numerical simulation of problems in mathematical physics: for example we can to mention here even such free software as Scilab or Python (x, y), but still, the main advantage to achieve maximum efficiency in interactive lecturing we should specify to the Mathematica software. Point is that the Mathematica provides many opportunities to display the results of the numerical solutions of equations of mathematical physics in the form of charts and graphs: 2D-plot, 3D-plot, contour plot, density plot, parametric plot, Video Graphics, 3D-Video Graphics. With new opportunities to export/import in this system (read numeric data, graphics and sounds in many standard formats - GIF, EPS, JPEG, AU, WAV, HDF) graphical results can be exported from Mathematica to formats of relevant kinds. The NDSolve is the powerful tool In Mathematica for numerical solution of the PDEs. The numerical method of lines is used here as a technique for solving partial differential equations via discretizing in all but one dimension, and then integrating the semi-discrete problem as a system of ordinary differential equations system or differential algebraic equations system. For the PDEs this method typically proves to be quite efficient. Currently, the only method implemented for spatial discretization is the Mathematica TensorProductGrid method, which uses discretization methods for one spatial dimension and uses an outer tensor product to derive methods for multiple spatial dimensions on rectangular regions. This specific method has its own set of options that you can use to control the grid selection process, what is very important property for best visualization. Below it is shown PDE (3) with initial and boundary values (4).

$$\frac{\partial}{\partial t} u(x, t) = \frac{1}{8} \frac{\partial^2}{\partial x^2} u(x, t), \quad (3)$$

$$u(x, 0) = 1.5 \cos(12x), \quad u(0, t) = 4 \sin(1.5t), \quad (4)$$

$$\frac{\partial}{\partial x} u(1, t) = 0.$$

After numerical solving, again using Mathematica 3D plotting commands, we can immediately create 3D graphics of surface plot of the solution in the specific range, as it is shown below, in the fig. 2.

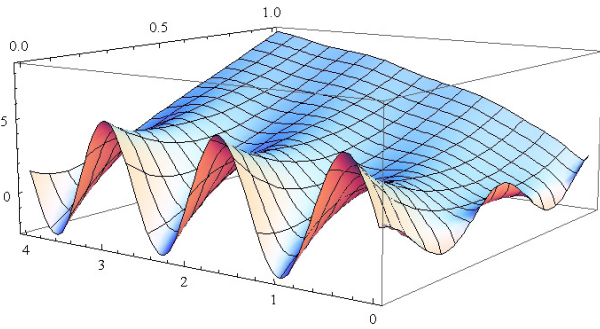


Fig. 2. 3D graphical visualization of the numeric PDE solution surface with boundary and initial values.

But more flexible tool with interactive visualization in Mathematica, for our point of view, it is Dynamic Visualizer package. This tool allows you to work with 3D images in real time. Graphic simulation of the various dynamic systems with Service Dynamic Visualizer - is a powerful and interesting addition to the visualization of Mathematica. The tool allows you to create three-dimensional images that previously were available only on expensive specialized workstations. All conversions can be performed and programmed interactively. Graphic objects created in Mathematica, via MathLink protocol passed to the Dynamic Visualizer, which can rotate, stretch and color. Dynamic Visualizer works with all standard graphics commands of Mathematica (ParametricPlot3D, Plot3D, Graphics3D, etc.). You can apply textures created in Mathematica objects to the tool - both static and animated too. Objects can be displayed or only using the edges (in the form of frames), or as a set of planes, or as a smooth bodies. In order to achieve a greater illusion of reality, Dynamic Visualizer allows you to choose illumination, light scattering, the spectral reflectance and transparency in real time. Below we present sequence of PDEs simulation examples via Dynamic Visualizer technology in CDF Wolfram Player, which are available from the Wolfram Research official site [6]. We should emphasize that for education purposes in mathematical physics the lecturers can get maximum advantage using this site. There are much interactive examples via CDF Wolfram Player.

The following example [7] simulates the Laplace's Equation on a Square in two dimensions is (5):

$$\frac{\partial^2}{\partial x^2} u(x, y) + \frac{\partial^2}{\partial y^2} u(x, y) = 0. \quad (5)$$

In the Fig.3 we can see that with use of the push-button switches on the model it is possible simulate different kinds of boundary conditions: linear, quadratic and cubic.

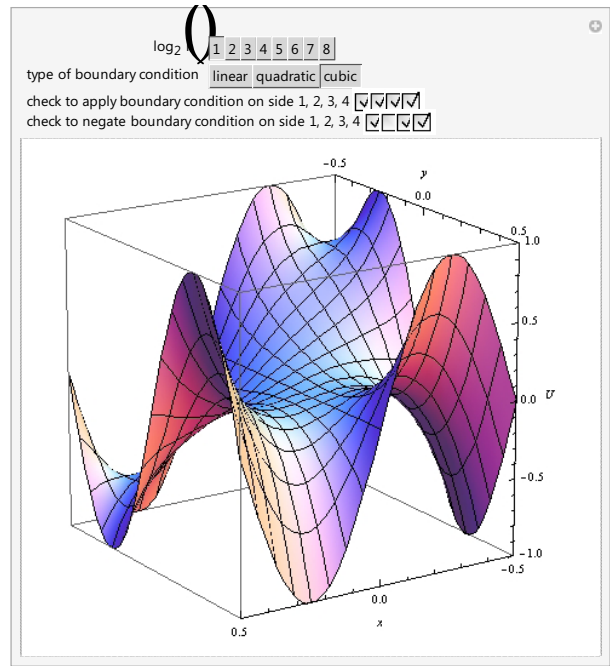


Fig.3. Laplace's Equation on a Square

At last, the following example, Fig. 4 is related with simulation of nonlinear wave equation [8]. In interactive regime the lecturer can explore different kind of nonlinear wave equations: from the ordinary linear wave equation to the sine-Gordon equation and Wolfram's equation.

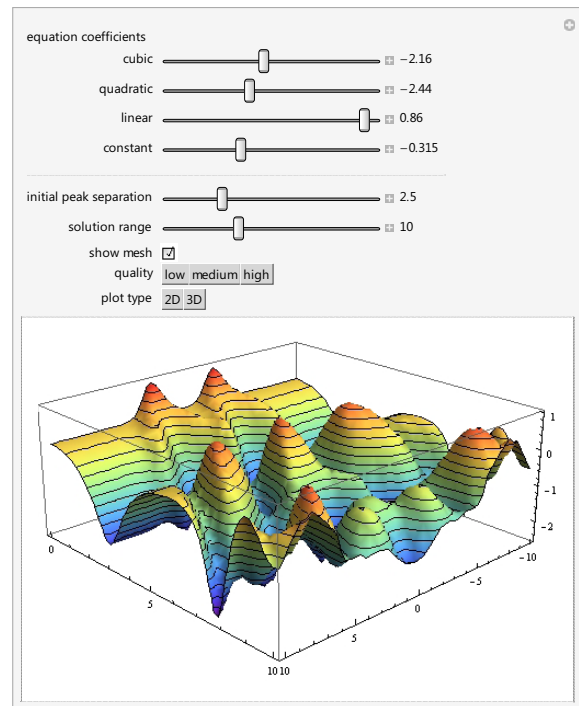


Fig.4. Nonlinear Wave Equation Explorer

IV. VISUALIZATION OF NUMERICAL SOLUTION OF PROBLEMS OF MATHEMATICAL PHYSICS BY FINITE ELEMENT METHOD

We definitely can say that the finite element method (FEM) is general numerical method in all scientific and engineering fields. It is well known that so many CAD systems have been created on the basis of this numerical method. For now day the researchers sufficient to know theoretical basis of this method and they should to manage with GUI of the CAD system for constructing the solution or designing of the researched physical object. Any way even in academic fields this method has big advantages. We definitely can say that the finite element method (FEM) is general numerical method in all scientific and engineering fields. It is well known that so many CAD systems have been created on the basis of this numerical method. Now researcher sufficiently to know theoretical basis of this method and he should manage with GUI of the CAD system for constructing the solution or designing of the researched physical object. As for interactive education of mathematical physics via computer visualization of results of computations on the basis FEM, then we have two approaches: - programming and developing original packages, - use CAD systems or tools. Let as consider first approach the original program package in Mathematica system which was developed by author in [9]. In this site author has described computational methods on the basis FEM, has presented developed by him the source program file in Mathematica and exe file via CDF Mathematica Player. Seven types of plane stress-strain problems have been considered here, Fig. 5. To all plane objects have applied different type of outside forces. In result you can visualize the output solutions in kind of FEM grid on the deformable bodies or visualize stress-strain behavior of deformable bodies in color. Seven type of plane stress-strain problems have been considered here. To all the constructively plane objects have been applied different type of outside forces. In result you can visualize the output solutions in kind of FEM grid on the deformable bodies or visualize stress-strain behavior of deformable bodies in color.

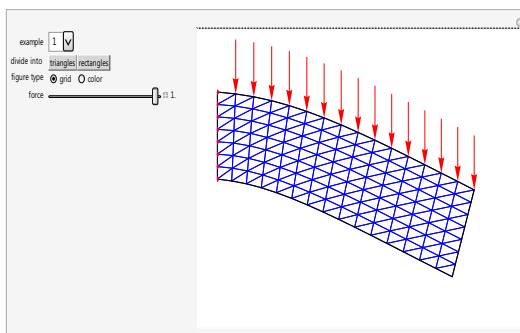


Fig.5. Stress-Strain Analysis by the Finite Element Method

Problems of mathematical physics those are the presented in Chapter II via interactively simulated in the CDF Mathematica Player, in each case they requires individual

approach to numerical methods and to techniques of programming. This means that problem of mathematical physics with specific geometric parameters, physical properties, with different boundary and initial conditions often requires of developing a unique numerical algorithms and programming methods. For example, the problem which is solved via FEM [9] has a complex source program in the language of Mathematica with more than a hundred lines. Therefore in order to solve specific problems the appropriate program source is need significant changes. Only in exceptional cases for a limited number of tasks in the CDF Mathematica Player you can develop an interactive visualization tool or use interactive examples, as a universal tool from the site <http://demonstrations.wolfram.com/topics.html?PhysicalSciences>. But there is another more rational way - is the use of systems modeling of physical problems, using FEM as a numerical tool but with elements of CAD technology. In the first place, we think that that the Matlab software has such features: Partial Differential Equation Toolbox with FEM as universal numerical method. The only disadvantage of this tool lies in the fact that the numerical algorithm for solving FEM constructed for PDEs with two variables. In order to solve problems in which the geometrical parameters of the physical object belong to the 3D, need to resort to special transformations, which are not applicable as an universal technique (e.g. when the problem is such that variation with third axis is negligible, all third axis - derivatives drop out and the 2-D equation has exactly the same units and coefficients as in 3-D). Any way this tool is much more acceptable for our purpose. Because via this tool you can simulate the elliptic and parabolic equations which are used for modeling: steady and unsteady heat transfer in solids, flows in porous media and diffusion problems, electrostatics of dielectric and conductive media, potential flow. The hyperbolic equation is used for: transient and harmonic wave propagation in acoustics and electromagnetics, transverse motions of membranes. The eigenvalue problems are used for: determining natural vibration states in membranes and structural mechanics problems. As an example, below in the Fig. 6, Fig. 7 and Fig. 8, are shown the graphical visualizations of results of computation of several problems in mathematical physics, which we have solved in Matlab via Partial Differential Equation Toolbox. The first problem is the current density between three metallic conductors (the physical model for this problem consists of the Laplace equation). To graphically visualize the current density need choose plot parameters in PDE Toolbox GUI, (in the Fig. 6 it is shown geometrical form of the plain body according to FEM discretization. In the Fig. 7 you can see colored graphics of contour lines and arrows distribution of current density along geometrical shape of the body. The following example is related to the stress-strain analysis of the flat bodies which have irregular, discrete shapes. The plate has the shape of isosceles trapezoid with a triangular hole. One side of the isosceles trapezoid is rigidly clamped, the others edges of the plate are free of outside forces and of supports. Two concentrated forces, which are

parallel to the base of the trapezoid, they are applied to the free vertices of the trapezoid. Such problems of mathematical physics related to plane problems of elasticity theory, and they can usually be solved analytically, using the functions of a complex variable. The results of the solution presented in Fig. 8, where the left one shows us the deformed plate state with deformed mesh of finite elements, thin lines shows the initial shape of the object. On the left in Fig. 8, we can visualize the color image of the deformed plate with a concentration of stress. To solve these problems analytically lecturer may spend considerable time, and the question about interactivity teaching would remain open. However in the Matlab via using Partial Differential Equation Toolbox we were able to solve this problem for a few minutes.

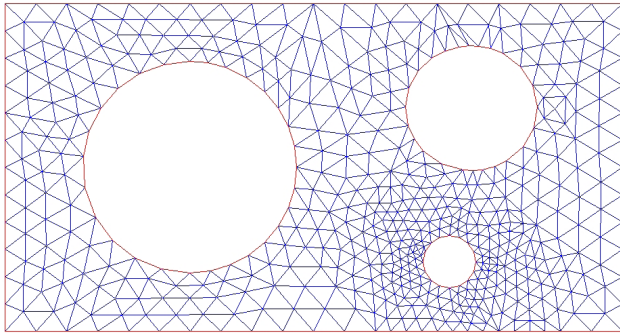


Fig.. 6. The current density between three metallic conductors with deformed mesh of finite elements

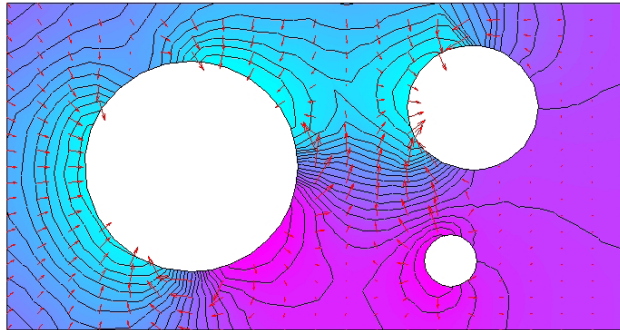


Fig.. 7. The current density between three metallic conductors (colored contour lines and arrows distribution of current density along geometrical shape).

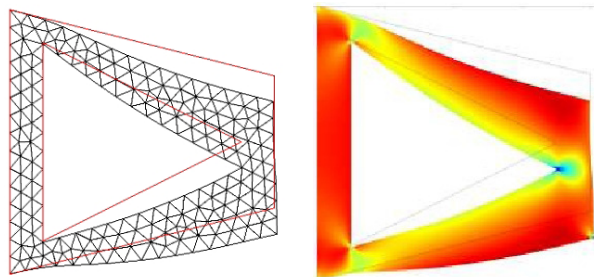


Fig.. 8. Flat stress-strain state of the trapezoidal plate with a hole in the form of a triangle.

CONCLUSIONS

In this paper we have attempted to uncover the opportunities that create application of computer-based visualization using advanced mathematical software to improve the quality of interactive teaching of students in courses of mathematical physics. We again emphasize that there are a number of computer programs of mathematics as a commercial and free, which have sufficient level of computer visualization of 2D, 3D graphics and graphical animation techniques. But the experience of teaching in this area is crucial. Therefore, we believe that the practical issues and analysis of critical problems, which we have discussed in the article, will make a contribution to improve the quality of teaching interactive courses of mathematical physics. We hope that our experience will be beneficial for lecturers of other related disciplines associated with PDEs and which use computer visualization techniques for improving the quality of interactive teaching.

ACKNOWLEDGMENT

We would like to thank the all referees which carefully have read the original manuscript, and in result they gave us theirs valuable comments, advises and suggestions. For our mind, theirs critical activity significantly has promoted to improving of the presentation of this paper.

REFERENCES

- [1] Richard S. Palais. (June-July 1999), "The Visualization of Mathematics: Towards a Mathematical Exploratorium", *Notices of the American Mathematical Society* 46 (6): 647–658.
- [2] F.François, G.Benjamin, B.Guillaume, "Sparse Meshless Models of Complex Deformable Solids," SIGGRAPH 2011 papers on the web, http://www-ljk.imag.fr/Publications/Basilic/com.lmc.publi.PUBLI_Article@12f67a0f733_189e6fb/main.pdf.
- [3] http://en.wikipedia.org/Comparison_of_computer_algebra_systems.
- [4] http://en.wikipedia.org/wiki/List_of_information_graphics_software.
- [5] http://en.wikipedia.org/wiki/Scientific_visualization
- [6] <http://www.wolfram.com/solutions/education/higher-education/>.
- [7] David von Seggern, "Laplace's Equation on a Square" the Wolfram Demonstrations Project, <http://demonstrations.wolfram.com/LaplacesEquationOnASquare/>.
- [8] Stephen Wolfram. "Nonlinear wave explorer". the Wolfram Demonstrations Project <http://demonstrations.wolfram.com/NonlinearWaveEquationExplorer/>.
- [9] Yuncong Ma. "Stress-Strain Analysis by the Finite Element Method", the Wolfram Demonstrations Project, <http://demonstrations.wolfram.com/StressStrainAnalysisByTheFiniteElementMethod/>, Published: December 5, 2011

A Web Application Tamper Proof Method Based on Text and Image Watermarking

Zetao Jiang

School of Information Engineering
Nanchang Hangkong University
Jiangxi, China
zetaojiang@126.com

Hongwu Zhang

School of Information Engineering
Nanchang Hangkong University
Jiangxi, China
zzhcwl@yahoo.cn

Abstract—With the rapid development of Internet, web attacking, as well as important pages tampering, again and again take place every day, webpage security is paid more and more attention. This paper presents a novel web application tamper proof method based on text watermarking and image watermarking, this method generate grayscale image with ASCII code, then make use of dual chaotic map in Encryption and decryption process as well as in process of embedding and detection of watermark. The results of experiment show that the method has relatively high reliability, and has application value to prevent webpage from tampering.

Keywords—Web Application; Text Watermarking; Image Watermarking; ASCII Code; Chaotic Maps

I. INTRODUCTION

With the rapid development of Internet, the influence of network information on social life is increasing every day; risk of being attacked is also growing. Therefore, protection of website information is paid more and more attention. Detection and prevention of tampered webpage is one of important content of network and information security protection. In this paper, fragile digital watermarking technology is very sensitive to documents changes, namely watermark would be destroyed as soon as documents are changed, so this characteristic of watermark could be applied to the webpage tamper proof system. In order to embed watermark information into web pages in traditional webpage watermarking scheme, there are some methods including insensitiveness of upper and lower letters in html tags, white space inserting, order of tag attributes and so on. But these methods make the code non-standard and change the size of webpage. Webpage, a kind of text document, contains less redundant information so that the hiding capacity is less. Therefore, this paper presents that generate grayscale image by ASCII code, then encryption and decryption as well as embedding and detecting of watermark with dual chaotic maps, in order to preventing web from tampering.

II. WEBPAGE WATERMARKING SCHEME BASED ON TEXT AND IMAGE WATERMARKING TECHNIQUES

Research about digital watermark has been widely performed, especially embed watermark into the image data [1-3]. However, steganography into the text is usually not preferred because of the difficulty in finding redundant bits in a

text document. To embedding information into a text document, its characteristics must be changed first. This characteristic can be either text format or characteristics of the characters. But the problem is that when a small change has been made to the text document it will be visible by intruders or attackers.

In this scheme, on the control client side, we generate watermark by using MD5 hash function. Then generate a grayscale image from webpage by making use of ASCII code and embed watermark into this image. There are many redundant data in this generated image, so the embedding capacity is become large. Finally, hiding this image embedded watermark into webpage. On the web server side, detect the watermark.

A. Watermark Generating

Two widely used hash functions could be used to generate message digest value; the two algorithms are MD5 and SHA-1. If these algorithms meet the requirement that be very sensitive to minor changes of original information, they would play a good role in application of digital watermarking. There are many papers which have make analysis and comparison about watermark generating algorithms [4-6]. For avoiding collisions of watermarks and making implement of algorithms be easy and fast, this scheme take MD5 as our watermark generating algorithm.

B. Generating of Grayscale Image

ASCII is short for American Standard Code for Information Interchange, letters and various characters are stored in computer in accordance with specific binary-coded rules. ISO took ASCII Code as international standard, and it is called ISO 646 Standard.

In ASCII Code, one byte on behalf an English character and the leftmost bit is 0. It can represent 128 characters, these codes's numbers range from 0 to 127 in decimal. A Chinese character consists of two bytes, the first bit of every bite is 1 and their numbers both range from 161 to 254 in decimal.

Grey level of every grayscale image is between 0 and 255, so we can take ASCII codes of English and Chinese characters in web pages as grey level. The height of image is equal to lines of webpage and it's width is equal to the length of the longest line, grey level of other white space is 255.

Identify applicable sponsor/s here. (*sponsors*)

The height and width are calculated as follows:

$$\text{Image height} = \text{No. of lines in webpage} \quad (1)$$

$$\text{Image width} = \text{Max}(L[i]); 0 \leq i < \text{No. of lines} \quad (2)$$

Grey level is get as follows:

$$f(x, y) \begin{cases} ASCII \\ 255 \end{cases}; 0 \leq x < \text{width}, 0 \leq y < \text{height} \quad (3)$$

C. Watermark Embedding

1) Embedding Algorithm of Image Based on Dual Chaotic Maps:

Chaotic system is a highly complex non-linear dynamic system, it is extremely sensitive to initial conditions and chaotic parameter, and the sequences generated by chaotic system have a good pseudo-random features, relevance, non-periodicity and complexity. Chaotic system is determined, the same initial conditions and system parameter generate the same sequence.

This algorithm uses Logistic map and Henon map [7-9]:

a) Henon Map

$$\begin{cases} x_{k+1} = 1 + y_k - ax_k^2 \\ y_{k+1} = bx_k \end{cases} \quad (4)$$

While $(1.05 < a < 1.8, b = 0.3)$, system produces chaotic phenomenon. When get x_{k+1} and y_{k+1} , calculate $Row = \lfloor Height * (|x_{k+1}| - 1) \rfloor$ and $Col = \lfloor Width * (|y_{k+1}| - 1) \rfloor$.

“Row” and “Col” represent location where watermark could be embedded into. “Height” and “Width” stand for image’s height and width.

b) Logistic Map

$$x_{n+1} = \mu x_n (1 - x_n) \quad (5)$$

$$x_n \in (0, 1), \mu \in (0, 4]$$

While $(3.5699456 < \mu \leq 4)$, system produces chaotic phenomenon. If $(x_{n+1} \leq 0.5)$, we get 0, else if $(x_{n+1} > 0.5)$, we get 1; then we can get a two-valued sequence.

c) Implement of Algorithm

Henon is a two-dimensional map, it can afford large key space and high security. Therefore, this map can be used to find out location where watermarks can be embedded into. Logistic map’s iterative equation is simple, and it’s parameter is only one, so it’s computing speed is very fast. We can use Logistic map to generate pseudo-random binary sequences for encrypting watermark.

Steps of algorithm as follows:

- Choose two initial values: x_0 and y_0 , two parameters: a and b , then fourth formula is computed by iterative

operation for several times to generate a pair of numbers: x_n and y_n ;

- Calculate $M = x_n * y_n$ and $x_0' = |M| - [|M|]$, take x_0' as initial value of Logistic map, and choose a parameter: μ . Then fifth formula is computed by iterative operation several times to ensure the map produces chaotic phenomenon;
- Webpage file is computed by MD5 algorithm to generate a 128 bits two-valued sequence and get another 128 bits two-valued sequence in Logistic map for encrypting watermark, the two 128 bits sequence are then computed by XOR operation to generate a new 128 bits sequence;
- Searching for the location where watermark could be embedded into in upper method, if the grey level of this location is 255, we can embed watermark at here;
- Converting watermark to grey level, 255 for 0 and 254 for 1; keep searching until the 128 bits sequence are embedded into image.

2) Hiding of Image Embedded Watermark:

In HTML, code of inserting of image as follow:

```
</img>
```

There are many attributes in start tag, eg: height, width, border and so on.

In order to hide grayscale image embedded watermark in webpage, we can write the code as follow:

```
</img>
```

Part of codes in webpage hidid grayscale image as follows;

```
</script>
</body>
</html>
<!-- syc92.search.sk1.yahoo.com compressed/chunked Sat Mar 31 19:30:55 PDT 2012 -->
</img>
```

D. Detecting of Watermarks

On the web server side, when the server receives user’s request to visit the website, detection process firstly gets the webpage which user request to visit, then extracts the image embedded watermark from the webpage. Secondly, process detects out encrypted watermark from the image and decrypts the watermark. Thirdly, process generate new watermark of webpage by MD5 operation. Finally, compares old watermark with new watermark. If the two watermarks are the same, server response to user’s request, if not, start alarm and recovery thread.

III. EXPERIMENTAL RESULT AND ANALYSIS

In this paper, we use VC++6.0 to implement the proposed scheme. We use the homepage of Yahoo Search as tested webpage and pictures are saved as a png format. The codes of the original web page can be seen in figure 1.

Figure 2 is the grayscale image generated from web page in figure 1. Figure 3 is the image watermarked. We can see that there is no visual difference between after watermark embedding.

Figure 4 and figure 5 are the web pages that users can see in web browser before and after the image is hid into the web page. We can see that there is no visual difference between after image hiding into.

A. Generating of Grayscale Image

The code of original webpage and grayscale image as follows:



Figure1 Code of Original Webpage

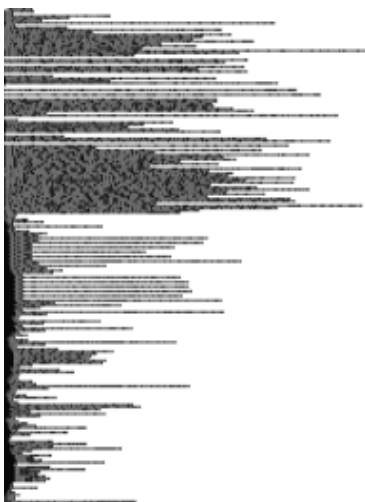


Figure2 Grayscale Image

B. Embedding of Watermark

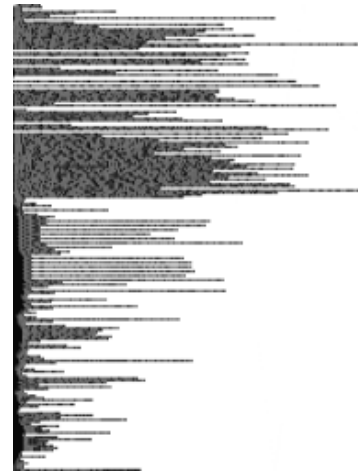


Figure3 Image Embedded Watermark

C. Inserting and Hiding of Images

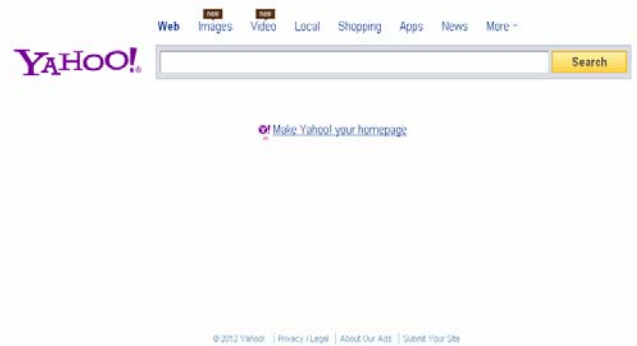


Figure4 Original Webpage

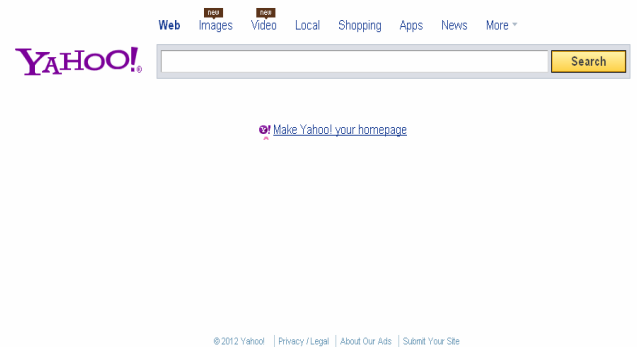


Figure 5 Hiding Image in Webpage

IV. CONCLUSIONS

This paper presents a novel web application tamper proof method based on text watermarking and image watermarking, this method uses MD5 algorithm to generate webpage watermark. Attacker maybe forged fake webpage which have the same watermark with original webpage by making collision of hash value [10]. Therefore, we make use of Logistic map for encrypting watermark. The innovation of this paper is to take

ASCII code of letters as grey level for generating image, which effectively increases the embedding capacity of watermark. Then make use of Henon map for find out location where could be embedded watermark into. Finally we hiding the image into webpage. The embedding algorithm has a good transparence in this method.

Results of experiment prove that both generating of image and embedding and detecting of watermark have a high degree of operational efficiency. Because chaotic system have a good randomness, confidentiality, volume keys and it's key is easy to replace, this method has a high degree of security.

REFERENCES

- [1] Jaseena K.U., Anita John. Text Watermarking using Combined Image and Text for Authentication and Protection, International Journal of Computer Applications, April 2011, Volume 20, No.4:8-13.
- [2] Cheng Nian-sheng, Study on Image Watermarking Algorithm Based on Chaotic Scrambling DCT, Computer Simulation, 2011, 28-6:288-291..
- [3] Vineeta Khemchandani, Prof G.N.Purohit. Information Security and Sender's Rights Protection through Embedded Public Key Signature, International Journal of Computer Science Issues, May 2010, Vol. 7, Issue 3, No 9:27-34..
- [4] C Wu, C Chang, S Yang, An Efficient Fragile Watermarking for Web Pages Tamper-Proof, Advances in Web and Network Technologies, and Information Management, 2007:654-663.
- [5] Wei Ruan-jie, Lu Hong-tao, A Novel Web Application Tamper Proof Method, Information Security, 2010, 71-73.
- [6] Peng Sun, Hongtao Lu, Two efficient fragile web page watermarking schemes, 2009 Fifth International Conference on Information Assurance and Security, doi:10.1109/IAS.2009.27.
- [7] Li Li-zong, Gao Tie-gang, Gu Qiao-lun, Image reversible data hiding algorithm based on chaotic system, Computer Engineering and Design, 2011, 32-12:4137-4142..
- [8] Li En, Wu Min, Xiong Yong-hua, Design and Application of Encryption Algorithm Based on Double Chaos Map, Application Research of Computers, 2009, 26-4:1512-1514
- [9] Li Li-zong, Gao Tie-gang, Chen Rong etc, Research and Application of Information Hiding Based on Chaos in Web, Chinese Journal of Scientific Instrument, 2008, 29-4:611-614.
- [10] Xiaoyun Wang, Hongbo Yu. How to Break MD5 and Other Hash Functions. Advances in Cryptology-Eurocrypt 05, LNCS3494:1-18, 2005

Interactive Systems For Sign Language Learning

Iurii Krak

Dept. of Intellectual Information Technologies
V.M. Glushkov Institute of Cybernetics
Kyiv, Ukraine
yuri.krak@gmail.com

Iurii Kryvonos

Dept. of Intellectual Information Technologies
V.M. Glushkov Institute of Cybernetics
Kyiv, Ukraine
aik@public.icyb.kiev.ua

Waldemar Wojcik

Dept. Electrical Engineering and Informatics
Lublin University of Technology
Lublin, Poland
waldemar.wojcik@pollub.pl

Abstract— In the article the problems of communication of deaf people uses sign language are consider. An analysis of sign language information transfer which includes human hands, body, fingers movements, change of mimicry and emotions on human face is brought. Conception is developed and new information technology is proposed for sign language modeling on the base of human spatial model. For the transmission of movements of the real human-informant of sign language on a spatial model technology of motion capture is used. For dactyl alphabet modeling technology that uses a three-dimensional model of a hand based on a informational-parametric model has been developed. Efficiency of the developed information technology is shown on realization of Ukrainian sign language. The proposed approach is carried by universal character and can be used for the modeling of other sign languages.

Keywords- deaf people, design, fingerspelling alphabet, sign language, information technology

I. INTRODUCTION

Modern progress in computing and creation of new methods for data representation, storage and organization makes it possible to innovate and create new technologies for sign language learning [1]. As to marked in [2]: "...Sign language interpretation services should also be provided to facilitate the communication between deaf persons and others...". From the practical point of view, teaching systems of sign language that use 3D human model are very promising. Since the information is transmitted by the mean of arms movements, mimics and articulation it is necessary to research the process of construction of a sign language sentence as well as the synthesis of the elements in order to get a good understanding of the subject. The problem of description of human movements is complex enough with a high percentage of fuzzy knowledge about human body and its physiology [3]. Thus it is important to conduct a research of the process of how the gestures are formed from the viewpoint of formalization for the problem of modeling using 3D model and for the problems of gestures analysis and synthesis

The sign language is the primary mean for communication for deaf people, has national traits (e.g. English and French [4,5], Russian [6] , Polish, Ukrainian [1] etc.). And there are two sign languages, different in grammar and sets of signs, used by the deaf:

- a common sign language, which is typically used in an everyday communication; its grammar differs largely from the spoken language;

- a calculative sign language, which is used for official communication; it has some aspects of the common one as well as fingerspelling; it does not have its own grammar and follows the natural spoken language.

Comprehension of a natural language by watching lips, "lipsreading", is an important skill for a deaf, because common people generally do not know the sign language. With regard to this fact, the solution capable of reading the text being pronounced by watching lips can be regarded as an alternative technology for facilitation of communication with people having difficulties in hearing or seeing. Besides, the visual-based speech recognition is an additional independent source of information for the problem of speech recognition itself, and can be used for improvement. The synthesized equivalent of the text received by lips reading will facilitate the communication for people having difficulties seeing.

In the given paper authors argue on the necessity to create a complex informational technology aimed to facilitate the non-verbal communication between common people and deaf people, people having hard difficulties in hearing or having damaged hearing. Will note that on the basis of the offered technology can be realized other sign language.

Therefore, the following problems statement has been formed: to create of computer-aided informational technologies for communication with deaf people, with intent for implementation, that would provide the following features:

- creation of a system for sign language fingerspelling units (dactyls) modeling based on 3D model of human palm;

- gesture synthesis for the common sign language and calculative sign language on a 3D human model;
- representation on a spatial (3D) human model the pronunciations process with regard to emotional components;
- lipsreading analysis and modeling.

II. MODELING OF A 3D MODEL OF A HAND AND ANIMATION OF FINGERSPELLING PROCESS

To teach fingerspelling, technology that uses a three-dimensional model of a hand based on a informational-parametric model has been developed. The technology allows observing hand from different viewpoint during the learning process, show sequence of letters etc. The main window is shown on a Fig. 1,

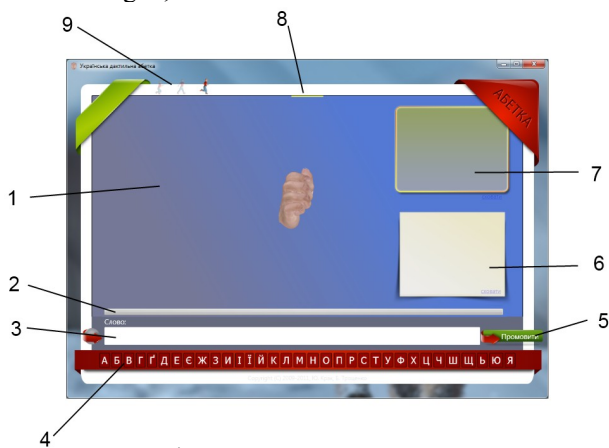


Figure 1. The main window of a program for fingerspelling alphabet modelling.

At this point the numbers mean: 1 – area of displaying fingerspelling alphabet; 2 – panel of displaying playback progress of letters or words; 3 – input panel for words; 4 – list of letters; 5 – button «spell», the process of fingerspelling of input word begins when the button is clicked; 6 – panel to display the verbal description of a hand configuration that correspond to the current displayed letter; 7 – panel to display written letter and a picture that correspond to the current displayed letter; 8 – indicator of a location of a hand rotation; 9 – define the pace of fingerspelling.

The main features of the program:

- Changing of a view angel. The use of a three-dimensional modeling enables the possibility to examine hand model from different viewpoints. That would be impossible using video materials. The range of changing an angle of viewpoint vary to 80° right/left;
- Presence of pictures which are associated with particular letter for whole alphabet (see Fig. 2, for example) (this panel can be hidden and shown back);
- Verbal description of hand configuration which is shown;



Figure 2. Examples of fingerspelling letters “O” and “X”, respectively.

- Presentation of dactyls is performed by selecting a particular letter from a list using mouse or by pressing letter-button on a keyboard. If user wants to repeat, press space. This feature allows to implement interactive learning process, when thee right arm (trained) is in the free position and the left realize interactions with the program;
- Changing the pace of the animation. Three pace modes (slow, medium, and fast) are implemented in the program for the different needs of the learning process (repetition after the model, recognition of the foregoing, etc.);
- Fingerspelling of a word. This feature allows entering words into input panel and observing process of fingerspelling of a word. That allows not only learn separate letters but also learn how to spell whole words;
- Verification mode. The program has a feature to “hide” the panel of the verbal description of a hand configuration and the panel with written letter and a picture. That allows conducting examination of knowledge displayed letter (hand configuration). Based on this technology, training programs for any one-handed fingerspelling alphabet can be created. There are currently developed programs for Ukrainian, Russian, Polish, Azerbaijan and American fingerspelling alphabets.

III. INFORMATIONAL TECHNOLOGY FOR SIGN LANGUAGE MODELLING AND LEARNING

The general approach for sign language uses for communication with deaf people is shown on a Fig. 3.

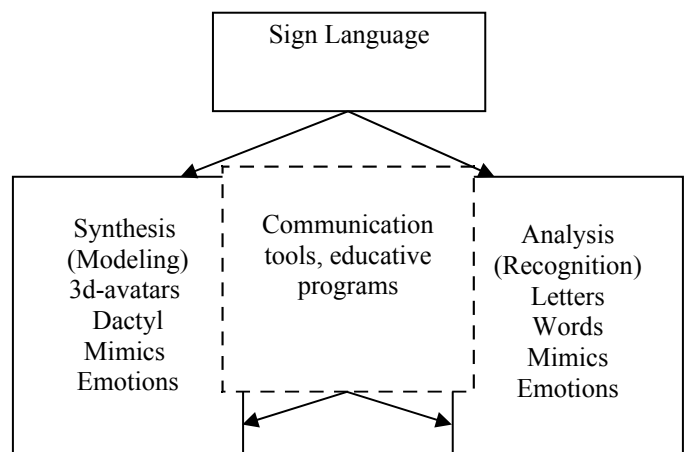


Figure 3. General scheme of the concept for communication with deaf people.

The complex information technology will provide the following features:

- a module for translation of the normal text into the sign language (text-to-gesture); the module will provide pronunciation animation of a common and official sign languages by presenting the output on a 3D human model;
- mimics and animation (with regard to emotional components) during the pronunciation process;
- lipsreading module for recognition of the text being pronounced.

For the implementation of the suggested concept of computer-aided non-verbal communication, a series of research works have been made and the appropriate software has been developed. For the 3D sign language animation synthesis, the geometrical classes of vector-based gestures are described. These classes were formed using motion capture technology [7]. Motion capture is a technology for retrieving real-world 3D coordinates using multiple video streams recorded from different viewpoints. Then the coordinates are used to determine values in the mathematical model. The key frames are determined by using tracking technology [8].

For the storage of a gesture, the BVH file format was used. It allows the gesture to be applied on a virtual human (e.g. using Character Studio module in 3D studio MAX or using Poser software). The suggested implementation of motion capture technology contains the following stages:

- 1) A person showing a particular gesture is recorded full-face, from the left and right side views;
- 2) The video streams are processed: arms coordinates are detected and the motion is tracked;
- 3) Based on the position of arms obtained on the previous stage, the BVH is formed for further synthesis of 3D animation;
- 4) The BVH is applied on a virtual human for creation of the animation process (using Character Studio in 3D studio MAX or Poser).

For the input text preprocessing, the appropriate informational technology was created, which considers the stress location for each word, specifies its normalized word form; contains synonyms and idioms. The model is represented as a set of tables in a relational database along with a set of stored procedures which implement all the required functionality. For the implementation of visualization and pronunciation feature of a custom text, the appropriate synthesizer has been created. It allows creating the voice equivalent of a custom text using different voices and voice characteristics (volume, distance). The synthesized allows to visualize the process of pronunciation by showing 2-dimensional visemes as well as 3D ones.

For the complex verification of the suggested technology the appropriate software has been created (Fig. 4). It is used for translation of a custom text into the calculative sign language.

- 1) The software uses the following algorithm for the sign language synthesis: 1) a speech equivalent is synthesized for the input text; 2) the input text is parsed into words; 3) a

speech equivalent is synthesized for the input text; 4) the input text is parsed into words; 5) for each word its normalized form (infinitive) is found by performing a look up in the database; 6) for each normalized word form a gesture is looked up (represented as a sequence of movements); 7) in case the gesture is not found, the word will be shown using dactyl alphabet.

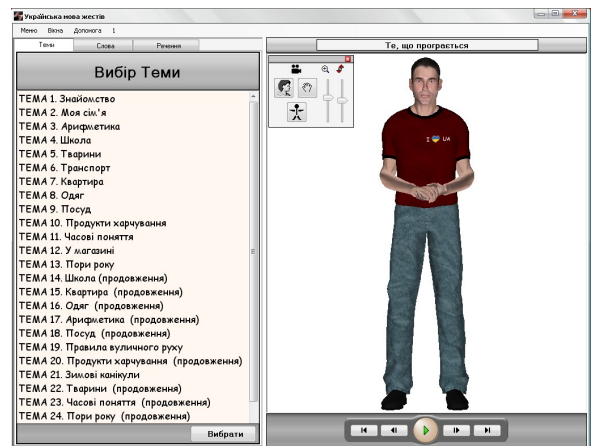


Figure 4. Computer system for sign language modelling and learning.

The 3D model displays the gesture accompanied by the speech synthesized.

CONCLUSIONS

The work suggests a complex of informational technologies for non-verbal communication with deaf people using sign language. Further development will be targeted at support for the full set of signs in the Ukrainian sign language. The great attention will be consideration on the use of the created systems for modeling Polish and other sign languages.

REFERENCES

- [1] Iu. G. Kryvonos, Iu. V. Krak, O. V. Barmak, ets. "Information technology for Ukrainian sign language modeling", Artificial Intelligence, no 3, 2009, pp. 186-197. (on Ukrainian).
- [2] Standard Rules on the Equalization of Opportunities for Persons with Disabilities. Appendix N 49 (A/48/49), article 14, pp. 292-306. Document 995_306, edited on 20th Dec 1993. <http://www.un.org/documents/ga/res/48/a48r096.htm>.
- [3] William C. Stokoe, Jr. Sign language structure: An outline of the visual communication systems of the american deaf. – Univ. of Buffalo, 1960.
- [4] C. Neidle, S. Sclaroff, V. Athitsos. SignStream™: A Tool for Linguistic and Computer Vision Research on Visual-Gestural Language Data, Boston University, Boston, Massachusetts, Vol.33, No.3, pp.311-320, 2001
- [5] IBM Research Demonstrates Innovative Speech to Sign Language Translation System, Extreme Blue programme HURSLEY, UK - 12 Sep 2007, IBM Media Relations <http://www-03.ibm.com/press/us/en/pressrelease/22316.wss>
- [6] A.L. Voskresenskii, G.K. Khakhalin. Multimediiniyi tolkovyi slovar russkogo gestovogo yazuka. Komputernaya lindvistika I Intellektualny tehnologii: Trudy mesgdunarodnoi konferencii «Dialog 2007» . - M.: Izd-vo RGGU, 2007. p.658. (in Russian).
- [7] A.Menache. Understanding Motion Capture for Computer Animation and Video Games, Morgan Kaufmann, 2000.
- [8] S. Avidan, "Support vector tracking" / Proc. IEEE Conference on Computer Vision and Pattern Recognition, Kauai, Hawaii, 8-14 December, vol. I, pp. 84–191, 2001

Quantum Concepts in Information Retrieval

M.Archuadze
Faculty of Exact and
Natural Sciences
I.Javakhishvili Tbilisi
State University
Tbilisi, Georgia
Maia.archuadze@tsu.ge

G.Besiashvili
Faculty of Exact and
Natural Sciences
I.Javakhishvili Tbilisi
State University
Tbilisi, Georgia
gela.besiashvili@tsu.ge

M.Khachidze
Faculty of Exact and
Natural Sciences
I.Javakhishvili Tbilisi
State University
Tbilisi, Georgia
manana.khachidze@tsu.ge

P.Kervalishvili
Faculty of Exact and
Natural Sciences
I.Javakhishvili Tbilisi
State University
Tbilisi, Georgia
kerval@global-erty.net

Abstract— Quantum information methods are successfully used with classic methods in information retrieval. This paper reviews the formation of quantum concepts, their presentation and storage and comparison methods for algorithms of information retrieval. We survey the representation of quantum concepts with qubits and their comparison methods via quantum logic gates. The paper discusses fuzzy concepts ranking, relevance evaluation and machine learning methods to improve information retrieval.

Keywords—quantum concepts, information retrieval, Semantic value, relevance

I. INTRODUCTION

Nowadays information processing is fundamentally studied with classical approaches; the latest improvements in this direction use existing explorations and no significant breakthroughs are observed. The explanation of such difficulties lies under the natural limitations to which we are already close enough. Our progress barely satisfies our needs, for we are reaching the edge of existing paradigms; consequently, we seek for novel approaches of information processing. Information processing methods based on quantum mechanical phenomenon is believed to be closer to nature, which promises to open a whole new world of opportunities. We see information processing based on quantum approaches as the future of information science [1, 2].

In quantum informatics, in difference from classic informatics, fundamental informational element is qubit. Via qubits, we can represent and process the information. Grover and Shore's algorithms are manipulated by qubits. The main goal of the researchers was to use the advantages of these algorithms and methods in classic computers [3,4].

II. QUANTUM LOGIC AND QUANTUM COMPUTATION TECHNOLOGIES

Selecting In different of classical logic in quantum logic the distribution low is not working (Heisenberg principle of uncertainty) [3].

The number of combined states of the system exponentially increased in case of quantum bits adding. This bringing us to the problem of quantum correlation estimation,

which is existing in case of quantum bits of integrated systems [5].

Example: Formation of superposition principle by using of Adamar's operator (WalshAdamar)

Superposition state and quantum states measuring effect as well is showing that physically exist the information fixed by user, which is located as quantum states observation in closed quantum system (until it will be influenced by external energy source).

This system is preserving close until the interaction with environment, and arisen main question is: how we could use the information located in superposition states? Within traditional formalism of quantum computation quantum operators are describing by equivalent matrix forms. Multiplication of operator's matrix with state's vectors is showing the action of operator on investigating system. For instance action of Adamar's matrix (H) on $|\psi\rangle = |0\rangle$ system should be written as:

$$H|\psi\rangle = H|0\rangle = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix} \begin{bmatrix} 1 \\ 0 \end{bmatrix} = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 \\ 1 \end{bmatrix} = \frac{1}{\sqrt{2}} (\begin{bmatrix} 1 \\ 0 \end{bmatrix} + \begin{bmatrix} 0 \\ 1 \end{bmatrix}) = \frac{1}{\sqrt{2}} (|0\rangle + |1\rangle),$$

and similarly

$$H|\psi\rangle = H|1\rangle = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix} \begin{bmatrix} 0 \\ 1 \end{bmatrix} = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 \\ -1 \end{bmatrix} = \frac{1}{\sqrt{2}} (\begin{bmatrix} 1 \\ 0 \end{bmatrix} + \begin{bmatrix} 0 \\ -1 \end{bmatrix}) = \frac{1}{\sqrt{2}} (|0\rangle - |1\rangle)$$

For computing processes the principle of superposition state importance becomes clearer if resulting superposition state will be interpreted as 2n classical trajectories will be computed by equivalent weights. All these create the opportunity to use the above mention weights as simplest elements for quantum parallel computing which should be performed by quantum computing hardware (machine). Following this idea superposition state is one of the first steps for quantum parallelism process organization.

Concepts (patterns) represented by bits transferred in qubits, for quantum concepts' comparison are used two qubits XOR (exclusive OR) Toffoli gates in Hilbert space.

A universal three-bit gate was identified by Toffoli [6] in 1981, called the Controlled-Controlled-NOT (or CC-NOT).

III. USING ONTOLOGY BASIS OF KNOWLEDGE REPRESENTATION

Humans are trying to describe the universe by natural languages, and this huge information is always developing. The task is: how to structure mankind knowledge and this is why the lexical ontology was born.

Formal view on lexical ontology brings us to semantic network, points of which are natural languages, while bonds represent the different dependencies existing in languages [7].

Ontology includes the abstract part, which represents the high level generalization of knowledge, and applied part performing the concrete pragmatic tasks of ontology. But knowledge is complete system about universe, which should be described by ontology methods and tools, and because of it within ontology the three level hierarchic was developed.

Three level hierarchy ontology is semantic network, points of which are the lexical signs of natural language's categories and notions, and they are bonding by forces based on semantic reasons. The lexical unit is calling concepts, and they could be interpreted by natural language vocabulary or by expert system. In ontology concept is represent by language notion, which has a knowledge content participating in formation of human's view about universe (internal and external).

Concepts and communications are created by expert system [8]. Following threelevel hierarchical ontology model communications might be formal (part, integer, element-multitude), and informal (with signs). According to logical interactions they could be binary or multiply and they are unite within ontology in one universal predict with A, B communication. There A and B represent the lexical concepts, which might be united in relevant clusters.

IV. CONCEPT FORMATION

Assume that we have multitude $S=\{S_i\}$, $i=\overline{1, n}$, and $\{S_i\}$ might be represent as natural multitude of numbers $N_A=\{1, 2, \dots\}$. Each element S_i of S multitude could receive value $j=1, 2, \dots, m$. For every real object of space $\{S_{ij}\}$, $i=\overline{1, n}$ $j=\overline{1, m}$ there is relevant trajectory $T(l)$, which is passed through one of the possible values of S_i . Multitude of all trajectories $T(l)$, where $l=mn$ might be represented by basic multitude $T(l+)$ $l=1, 2, \dots, k$ $k \leq mn$.

We should take one positive value $-\varepsilon$ relevant to $T(l+)$ trajectory, which represents certain real structure. We should call trajectories with $+\varepsilon$ value real (RT). For the same value $\{S_{ij}\}$ trajectories with $-\varepsilon$, should be called additional trajectories (AT), and trajectories having $+\varepsilon$ and $-\varepsilon$ estimations – generalized trajectories $T(lc)$.

The groups of generalized trajectories have one common basis within conceptual vector space. At this common basis generalized concept has relevant canonical form (CF), which $CF(C)=(C_1, C_2, \dots, C_n)$. In this conceptual space basis of which is $((l_1, l_2 \dots, l_n)$ presenting concept are computable. $\{S_{ij}\}$ - is

notion of equivalency and this gives a possibility to range trajectories by classes. If we consider semantic similarity of concepts it becomes possible to build hierarchical systems of ontological and semantical classes. At the same time we should considered the size of semantic similarities which is based on value of the distances in between them. These distances are determined as Euclid's sizes, Minkovsky or Hamming sizes.

If we will take A and B objects and their relevant concepts, semantical similarity size could be computed by: $Sim(C(A), C(B)) = \frac{2IC(dis(C(A), C(B)))}{IC(C(A)) + IC(C(B))}$, where IC – is Information Content, $(dis(C(A), C(B)))$ distance between concepts, $IC(C) = -\log P(C)$ – average value of informational entropy. $IC \geq 0$ is situated between 0 and 1; it is monotony and additive.

For comparison of similar concepts it is possible to use quantum methods, and particular Born's rule [9], which creates connection among mathematical description and experiment:

$$P(\Phi, \Psi) = \frac{|\langle \Phi | \Psi \rangle|^2}{\langle \Phi | \Phi \rangle \langle \Psi | \Psi \rangle}$$

$$P(\Phi, \Psi) = \frac{|\langle \Phi | \Psi \rangle|^2}{|\langle \Phi | \Phi \rangle|^2 + |\langle \Phi | \Psi \rangle|^2}$$

Where $\Phi \wedge \Psi$ is external (Grossman) multiplication of Φ and Ψ vectors.

If we consider S multitude of states as $S = \{S_1, \dots, S_N\}$ then states of the system and apparatus is described by transferring apparatus by n and m vectors:

$$|n\rangle = \begin{bmatrix} n_1 \\ \vdots \\ n_N \end{bmatrix} \quad \text{and} \quad |m\rangle = \begin{bmatrix} m_1 \\ \vdots \\ m_N \end{bmatrix}$$

Where n and m are natural numbers, S_i is common multiple of input.

Because of the symmetry these numbers are not observable; it is only possible to observe their invariant combinations. Because of the fact that scalar multiplication is invariant to its transferrable image, we can write:

$$P(m, n) = \frac{\sum_i m_i n_i}{\sum_i m_i^2 \sum_i n_i^2}$$

It is obvious that for non zero vectors with n and m natural components the image will be rational number with value more than zero. Therefore it is clear that destructive quantum interference phenomenon is principally impossible.

V. INCREASE OF INFORMATION SEARCH EFFECTIVENESS BY USAGE OF FUZZY CONCEPT METHODS

Task of information search and filtration n data bases and internet becomes very complicate in case of semantical properties of datas. During semantical search processes it is necessary to precisely describe document and its keywords [10].

Describing document by fuzzy model

Assume that data base includes n documents

$$D = \{d_1, d_2, \dots, d_n\}$$

I stage: **Determination of basic concepts**

User collects important for him multitude of basic notions. Assume that multitude of this kind concepts is:

$$T=\{t_1, t_2, \dots, t_m\}, \text{ where } t \text{ is:}$$

$$\frac{M(k_1)}{k_1} + \frac{M(k_2)}{k_2} + \dots + \frac{M(k_p)}{k_p}$$

$M(k_i)$ is function of k_i term of concept possession.

We can determine the degree of possession following two ways: first, when user giving the values to concepts himself, and the second – when indexes are determining automatically.

Semantic value of fuzzy concept for each document is illuminated by aggregative function $M_i(t)=A(k_1, k_2, \dots, k_p)$, which should be selected by minimization, maximization, average or another methods.

For instance, we used maximal value of the function, and our document could be described by following multitude of concepts:

$$d=\left\{\frac{M_i(t_1)}{t_1}, \frac{M_i(t_2)}{t_2}, \dots, \frac{M_i(t_m)}{t_m}\right\} \quad (1)$$

Elaboration of fuzzy thesaurus of request

Request should be prepared by the same form as document using formula (1).

In case of thesaurus preparation it is obviously the usage of the same concepts as in (1).

$$q=\left\{\frac{M_i(t_1)}{t_1}, \frac{M_i(t_2)}{t_2}, \dots, \frac{M_i(t_m)}{t_m}\right\}$$

Estimation of request relevance

Now it is necessary to underline that document classification and request preparation in dependence of cataloged data basis will not have the same estimation. Following that we should say (estimation) that requested documents will have fuzzy coefficient of relevance. Represent the multitude of documents, which are received for analysis after the request was satisfied. C represent the multitude of documents, which were requested, and $C = \{d | a_R(d) > 0, d \in D\}$.

$\alpha_R(d)$ - generalized coefficient of relevance is computing by formula:

$$\alpha_R(d) = \max(\min(M_i^d(d), M_i^q(d)))$$

Classification of request result in connection to relevance groups:

$$O = a_0 < \alpha_1 < \alpha_2 < \dots < \alpha_{u-1} < \alpha_u = 1$$

Using above mentioned parameters C_j could be represented as follows:

$$C = \{d | \alpha_R(d) = \max(\min(M_i^d(d), M_i^q(d)))\}$$

Size ordering model

Size ordering model includes the arrangement of selected documents according to requests degrees decrease. Within prepared list highly relevant documents have the high range, and they are situated at the head of list. In this case searching algorithms system is working on basis of relevance function computing, which is equal to parameter describing connection of searched document number to coming request. Therefore we have the process of computing of relevance of each investigated document according to request's content, and afterward preparation of list of documents also in accordance with their numbering.

For relevance computation different factors or community of factors could be used. But their main property should be the possibility of best selection of relevant documents from

irrelevant ones. The search systems are different to each other by factors, which are mostly similar. But factors computation rules are not the same for different systems. For this it is possible to use the simple assumption method or weight assumption or another more complicate method. The number of factors for these systems is not many, often not more than 15.

Last achievement for search technologies and tools creation demonstrates the systems with hundreds of factors, as well as their different combinations. In many cases machine learning method of algorithm optimization is positively influenced of search result, where addition of novel factors is not creating of any serious problem [11].

VI. MACHINE LEARNING

On the basis of learning parameters observations was elaborated the table in which factors could be estimated by numbers and by degrees. When we have into consideration the fuzzy concepts the fuzzy estimation might be used.

For instant, if we prepare the list of n documents in range of their relevance estimation and user is searching the interesting document from top to bottom, the size of estimation represents the probability that user will find in list the relevant result. According to this:

$$P_i = \sum_{i=1}^n P_i(\text{in the list}) * P_i(\text{relevant}),$$

Where P_i (in the list) is probability of searching of i document.

P_i (relevant) is probability of relevance of i document.

P_i (relevant) is also the numbered estimation of request relevance.

Calculation of probability of P_i (in the list) is based on the assumption that user is observing the table from top to bottom and stopped this process when he will found the relevant document or because of other reason. In this case:

$$P_i(\text{in the list}) = P_{i-1}(\text{in the list}) * (1 - P_{i-1}(\text{relevant}) * 1 - P(\text{stop})),$$

Where P_{i-1} (in the list) is probability that user will find (i-1) position.

$1 - P_i$ (relevant) is probability that user will not be satisfied by (i-1) position.

$1 - P_i$ (stop) is probability of that user will not stop by the reason independent from us. This parameter should be selected.

For relevance size ordering function computing could be used polinomic methods [12]. The type of polinoms should be selected by genetic algorithms. Coefficients in these formulas could be prepared using differential evolution algorithm.

In information retrieval problems, quantum concepts (patterns) are mainly used, for which Quantum arithmetic and quantum logic gates are used, that enables us to use Toffoli XOR gates for solving quantum concepts problems. But we haven't discussed the item that the storage of the quantum concept in base (memory) is rather difficult, because it changes from reading to storage. There are different ways to solve this problem; one of them is Probabilistic cloning

machine. For searching and comparing fuzzy quantum concepts, it is possible to use Humming measure. All these matters are the subject of the next papers.

REFERENCES

- [1] M. A. Nielsen and I. L. Chuang, "Quantum Computation and Quantum Information", Cambridge University Press, Cambridge (2000).
- [2] A. O. Pittenger, "An Introduction to Quantum Computing Algorithms", Birkhäuser, Boston (2000).
- [3] D. Aerts, M. Czachor, S. Sozzo. "Quantum Interaction Approach in Cognition, Artificial Intelligence and Robotics". arXiv:1104.3345v1 [cs.AI] 17 Apr 2011.
- [4] P.J.Kervalishvili."Quantum Information Science: Some Novel Views". Proceedings of the International Scientific Conference ICTMC-2010, Information and Communication Technologies – Theory and Practice. Nova Publishers, Series: Computer Science, Technology and Applications. ISBN: 978-1-61324-870-6, 2011. p. 9-17.
- [5] T. Toffoli, "Reversible Computing," Tech. Memo MIT/LCS/TM-151, MIT Lab. for Com. Sci. (1980).
- [6] Peter Witte, S'andor Dar'anyi. "Introducing Scalable Quantum Approaches in Language Representation". Quantum Interaction 5th International Symposium, QI 2011 Aberdeen, UK, June 26-29, 2011.
- [7] Jun Zhai, Yiduo Liang, Yi Yu and Jiatao Jiang. "Semantic Information Retrieval Based on Fuzzy Ontology for Electronic Commerce". School of Management, Dalian Maritime University, Dalian 116026, P. R. China, Journal of Software, Vol. 3., N9, 2008, pp. 20-29.
- [8] JM.Alonso, L.Magdalena. "A Conceptual Framework for Understanding a Fuzzy System". Proceedings of the Joint 2009 International Fuzzy Systems Association World Congress and 2009 European Society of Fuzzy Logic and Technology Conference, Lisbon, Portugal, July 20-24, 2009, pp. 119-124.
- [9] Christopher A. Fuchs. "Born's Rule as an Empirical Addition to Probabilistic Coherence" Quantum Interaction 5th International Symposium, QI 2011, Aberdeen, UK, June 26-29, 2011.
- [10] K. van Rijsbergen. "The Geometry of Information Retrieval". Cambridge, 2004.
- [11] M.Baziz, M.Boughanem, G. Pasi, H.Prade. "A fuzzy Set Approach to Concept-Based Information Retrieval". EUSFLAT – LFA, 2005.
- [12] E.Hüllermeier. "Fuzzy Sets in Machine Learning and Data Mining". Applied Soft Computing, Volume: 11, Issue: 2, Publisher: Elsevier, pp. 1493-150, 2011.

Decision Support System For Crisis Management Using Temporal Fuzzy Logic

Eng. Ammar Alnahhas
Faculty of Information Technology
Damascus University
Damascus - Syria
Email: a.nahhas@live.com

Assist.Prof. Bassel Alkhatib
Faculty of Information Technology
Damascus University
Damascus - Syria
Email: basselk@scs-net.org

Abstract—Crisis management involves a lot of factors and variables that have to be monitored and considered. Human judgment becomes far from optimal as these variables and factors increase in number. Thus, in these situations, we need the help of decision support system to help the decision making process. In this paper we observe the main decision making factors that should be considered in crisis management situation, we define the crisis as a substantial change in the value of regular environment variables and factors, as the environment is uncertain we propose to use a fuzzy set theory to represent the variables of the system, we show how we can use temporal logic to represent the temporal nature of crises, we present a combination of fuzzy and temporal inference techniques, and propose an algorithm to support the decision making, the algorithm is based on temporal simulation which make use of temporal relations represented by temporal statements, we finally show that the proposed model and algorithm can suite a lot of crisis management situations.

Keywords: crisis management, fuzzy logic, temporal logic.

I. INTRODUCTION

Crisis management can be defined as special measures taken to solve problems caused by a crisis and to minimize the damage to the organization [1], the crisis is an unstable or crucial time whose outcome will make a decisive difference to better or worse, we can observe how critical is the decision making process in the case of crisis.

Since the emergence of computer systems, many of them have been used by decision makers to support their decision in crisis situations, many technologies have been used, such as databases and spreadsheets. Nowadays as artificial intelligent has catch more interest of researchers; intelligent decision support systems have got more applications, they can be applied to the case of crisis management since the decision makers that deal with crisis situation need intelligent instruments like software tools capable to deal with questions related to the global perspective thinking [2], many researches have been involved in this field, most of them focus on a specific crisis situation, but some are generalized, such as [3].

Crisis has many characteristics [4], but we are interested in the following: 1) An unusual volume and intensity of events, 2) Sudden change in one or more basic system variables, 3) Change in the internal or external environment, 4) Uncertainty 5) surprise, these characteristics are almost common among

many crisis situations, so when a system design can model this characteristics, it can computationally support decision as it can model and simulate the environment changes. The system we present in this paper models the crisis management elements, we use fuzzy set theory to model uncertainty of system variables value that affects internal and external environment, temporal logic can model events, how they affects the variables, and how the variables interacts temporally, the resources usability is modeled in a linear logic like form, finally we use this models in a temporal algorithm that can simulate the environment variables value change and suggest possible actions, we show how the system can be used as a generic DSS for most crisis management cases.

II. FUZZY SETS AND ENVIRONMENT VARIABLES REPRESENTATION

Fuzzy set theory [5] was introduced to represent uncertainty, as apposite to ordinary set; where each element either completely belong to the set or don't, in the fuzzy set an element has a degree of membership, the fuzzy set is represented mathematically as a membership function which is a mapping from the universe X to the unit interval $[0,1]$. A fuzzy variable is a special variable whose values are fuzzy sets, each fuzzy variable has a set of discourse which is the crisp set that all the fuzzy set values of the variable is defined upon, each fuzzy variable has linguistic values that are linguistic names linked to a special values of the fuzzy variable. A fuzzy predicate has the shape $X \text{ is } V$ where X is a name of a fuzzy variable and V is a fuzzy set defined on the set of discourse of X , in fuzzy logic each predicate has a degree of truth, which is called a truth value, that is, each predicate may be partially true, and partially false.

Fuzzy variables are a good tool to model the environment variables in the case of crisis, as it is not possible to get a precise information about the state of the environment during a critical situation, for example in the case of fire, we can observe that the smoke is dense, but it is very difficult to measure the pressure of smoke accurately in this situation. So the predicate "Smoke is dense" can represent the state. Fuzzy representation provides another feature that can be exploited which is the fuzzy sets interaction, the two predicates Smoke is dense and smoke is moderate are semantically connected,

since the fuzzy sets "dense" and "moderate" are linked and can have some measure of similarity, this feature provides a good starting point of evaluating the current state of the environment. The environment state can be represented by variables values, so we can define the vector V where

$$V \in F(A_1) \times F(A_2) \times \dots \times F(A_n) \quad (1)$$

Where $F(A_n)$ is the set of all fuzzy sets on A_n , A_n is the set of discourse of variable n , this way we can represent the state of crisis where the definition of crisis is related to a substantial change of system variables. The goal of the system is to restore the normal variable values, which in turn can be represented as another vector of values defined in the same method. To evaluate the situation we should define a similarity function that maps two vectors representing two different states, the function

$$Vsim : (F(A_1) \times F(A_2) \times \dots \times F(A_n)) \times (F(A_1) \times F(A_2) \times \dots \times F(A_n)) \rightarrow [0, 1] \quad V(\bigcirc P, t) = V(P, t + 1) \quad (2)$$

$Vsim$ can be defined if we can find a similarity measure between two fuzzy sets S , the similarity Measure should satisfy the some conditions [6], In the case of crisis management the similarity is better not to be linear since the variable values are crucial, we can define the similarity as a function of Distance:

$$D(A, B) = \sum_{i=1}^n |A(x_i) - B(x_i)| \quad (3)$$

$$S(A, B) = e^{-\alpha L(A, B)} \quad (4)$$

Where α is a slope factor, its value determines how critical the situation is.

We can define the function $VSim$ by means of S

$$VSim(V1, V2) = Mini(S(A_i, B_i)) \quad (5)$$

the goal of the system can be defined using $VSim$, if we denote the current system state as C , and the target system state as T then the goal is

$$VSim(C, T) > Threshold \quad (6)$$

Where threshold value determines the significance of system goals.

III. TEMPORAL LOGIC AND TEMPORAL RELATIONS

In temporal logic [7], predicate truth changes over time, that is, a statement may be true now but not in the future, linear time temporal logic is one of the temporal logics where time is considered linear, where each time point has exactly one next point and one previous one, apparently the time model we use here is discrete, therefore the time point can be expressed as a natural number, and the relation $<$ greater than can be used to define the semantics of "next" and "previous".

We use temporal modalities to express time, predicates with no modalities are considered to be in the present, the \bigcirc means "in the next moment", \square means "always" and \diamond means "eventually", the modality U means "until" and it is a binary operator, in classical temporal logic "until" means that the

predicate precedes it must still true until the predicate follows it becomes true.

When using fuzzy predicates in temporal form, things are different because fuzzy predicate has a truth value which belong to the unit interval, Therefore the mixture of temporal logic and fuzzy representation can express the change of the truth value over time. In this case the meaning of the temporal modalities is different, in each time point a predicate P can have a truth value in the interval $[0,1]$, so $\square P$ is also fuzzy and can have a truth value interval $[0,1]$. Let the function V where

$$V : P \times N \rightarrow [0, 1] \quad (7)$$

$V(P, t)$ denotes the truth value of predicate P in time point t , the function V can model the change of the state variables over time, so it is important to predict V values in the future, the truth value of temporal modalities are defined as follows

$$V(\square P, t) = Min_{u \geq t}(V(P, u)) \quad (9)$$

$$V(\diamond P, t) = Max_{u \geq t}(V(P, u)) \quad (10)$$

The modality "until" is very useful to links variables in the crisis management situation, for example: the statement (*fire is strong*) U (*water_dump is high*) can be expressed, the value of this modality can be defined as

$$V(PUQ, t) = Min_{u \geq t}(Max(0, V(Q, u) - Min_{v \leq t}(V(P, v)))) \quad (11)$$

Back to the system goal in 6, the crisis management system should get to the goal in the future so we can rewrite the goal as

$$\diamond \square (Vsim(C, O) > Threshold) \quad (12)$$

Which mean that "eventually the system should be stable forever", we can easily observe that the temporal modalities can express the feature of function V , as they define the minimum, the maximum and possibly the value in the next moment of the function which can model the system state change.

IV. TEMPORAL EVENTS

Events can be defined in terms of system variables, each event can be linked to a variable, so that the event is the variable having a specific value, for example, we can define the event of "low power" as the variable "power" having the value "low", so the event is linked to the predicate "power is low", external events occur outside the system so they are not linked to variables. Formally we can say that the event $E(Var, Val)$ occurs when variable Var has a Value Val . We can easily observe that E has a degree of occurrence because variable are fuzzy and its value can be partially true, so we define the degree of occurrence as the truth value of the predicate denoting the event, if the function V can be used to express the degree of occurrence of event E in a time point t we can write

$$V(E(Val, Var), t) = V(Val \text{ is } Var, t) \quad (13)$$

The event defined above can be called *instantaneous event*, because it happens in a single time point, we can define some continuous events in the environment of crisis management, the event of the variable holding the same value for a while is one of them, we can define the event *for* which is $For : F(A) \times N \rightarrow [0, 1]$, $For(VarisVal, L)$ means the Variable Var has been having the value Val Since L time steps before now, the degree of occurrence of this event can be defined as

$$V(for(P, N), t) = Min_{t-n \leq u \leq t}(V(P, u)) \quad (14)$$

Another continuous event may be defined as the fact that some variable changes its value, the event $To : V \times F(A) \times F(A) \times N \rightarrow [0, 1]$, where $To(V, Val_1, Val_2, L)$ means that the variable V has changed its value from Val_1 to Val_2 in the last L time steps, the degree of occurrence can be defined as

$$V(To(v, a_1, a_2), t) = Min(V(v is a_1, t - n), V(v is a_2, t)) \quad (15)$$

V. RESOURCE CONSUMPTION REPRESENTATION

Managing resources is very important in the process of crisis management, they are important to perform actions, some resources are not reusable, and some have a temporal restriction for reusability. In our work we suggest to use the model of Linear logic [8] to represent resources usage, Linear logic is a special type of logic that is widely used to manage resources. In linear logic each atom represents a resource, and can be used once, when it is used it becomes unavailable and can't be used at the same time. Temporal linear logic has been used to represent resource usage over time (see [9]), in this logic \bigcirc means the resources can be used in the next time point, the \square modality means that resource can be used at any time in the future but exactly once, \diamond can represent future availability, it means that the resource is going to be available in the future but we are not sure exactly when, using those modality we can express resources usability and reusability using logical statements, for example, $\bigcirc \bigcirc \square A$ means that after two time steps you may use the resource A whenever you want but exactly once.

Implication can represent the temporal availability, for example, $A \multimap \bigcirc \bigcirc \square A$ says that when you use the resource A it will be available again after two time steps. Resource consumption will be linked to actions as we will show in the next section.

VI. ACTIONS AND DECISIONS

Decisions are the main output of the system, decisions can be defined as an action that affects the values of system variables, either immediately or after a period of time, additionally some resources may be consumed.

Action may set a new value to environment variable, or may change current variable value, in the first case the Action can be defined as a triple (P, R, E) where P is the predicates describing new variables values, R is the set of resources needed, and E is a set of events, which the action should be taken upon occurrence.

We can represent the change of variable value as a shift in the fuzzy set that represents the value of the variable, shifting can be done using the formula:

$$A(x) = A(x + n) \quad (16)$$

Where n is related to the action and can be linked to a fuzzy set defined on the special variable "change", n then can be obtained as a defuzzification of this fuzzy set.

VII. INFERRING FUTURE VARIABLE VALUES

The goal of temporal relations is to find the future values of system variables, the temporal relation links variable values from past and present to other variables in the future, so the inference aims at finding the expected fuzzy set that represent the value of the variables in all future time points, this will help evaluating the future state of the environment, the temporal relation has one consequent and one antecedent each has a tense which is either past, present or future, and the inference rule can be found for each type of relations. When a relation links present to present, it has the formula $X is A \rightarrow Y is B$, here we can use the well-known fuzzy compositional rule of inference

$$\frac{X is A \rightarrow Y is B \quad X is A'}{Y is A' \circ R(A, B)} \quad (17)$$

When a relation links present to future, the formula is $X is A \rightarrow \alpha P$ where α is one of the temporal modalities $\bigcirc \square \diamond$, we can use the rule

$$\frac{\frac{X is A \rightarrow P}{P'} \quad X is A'}{X is A \rightarrow \alpha P} \quad X is A' \quad (18)$$

The relation may link the past events, which are the continuous events that occurs in a period of time that ends at present, to the present variable values, the rule formula is $E \rightarrow X is A$, in this situation the antecedent of the relation is not a fuzzy predicate of shape $X is V$, to use a standard fuzzy inference rule, we can consider the "true" fuzzy set which can be defined by membership function $true(x) = x$, we can then write the event E as $E is true$ to standardize the rule shape, now having the last relation and the event E with degree of occurrence equals to p we can infer A' where

$$A'(y) = Min(b, A(y)) \quad (19)$$

We can extend this rule to be applied to the relations that link past to the future, the form of the relation is $E \rightarrow \alpha P$ where α is one of the temporal modalities $\bigcirc \square \diamond$, if the event E having a degree of occurrence n is denoted as $E(n)$ we can use this rule

$$\frac{\frac{P \rightarrow QP(n)}{Q'} \quad P \rightarrow \alpha Q \quad P(n)}{\alpha Q'} \quad (20)$$

When the relation links the future to the future, it can be easily treated as the type of relation from present to future, as we can use the following temporal formula

$$\alpha A \rightarrow \alpha B \Leftrightarrow \alpha(A \rightarrow B) \quad (21)$$

when the relation has more than one antecedent, we can merge the results of rules using this rule [10]:

$$\frac{P_1 \rightarrow X \text{ is } Q_1 \quad P_2 \rightarrow X \text{ is } Q_2}{P_1 \wedge P_2 \rightarrow Q_1 \cap Q_2} \quad (22)$$

The "until" modality can be used as an inference tool, if we have the statement $P \text{ is } A \text{ U } Q \text{ is } B$ we can link the left hand side of u to the right hand side

$$\frac{P \text{ is } A \text{ U } Q \text{ is } B \quad \bigcirc(Q \text{ is } B')}{\bigcirc \square P \text{ is } A'} \quad (23)$$

Where $A'(x) = \text{Min}(1 - S(B, B'), A(x))$

Additionally, standard temporal inference rules can be used

$$\frac{P}{\square P} \quad \frac{\square P}{\diamond P} \quad \frac{\square P}{\bigcirc P} \quad (24)$$

VIII. MAKING DECISIONS

During the crisis management process, the decision maker should take actions and procedures that helps to get to the goal, our proposed algorithm evaluate the decision that may be taken by going into future and observe the effect of the decision on the system variable values. The main technique to get the future values of the variables is to use the temporal relations. The algorithm main process is to virtually advance the time index, the global working memory stores all temporal predicates that describes the future, we can easily update this memory when advancing the time index by the following substitutions:

$$\bigcirc P \Rightarrow P \quad (25)$$

$$\diamond(X \text{ is } S) \Rightarrow \diamond(X \text{ is } S) \text{ if } S \subseteq \text{Val}(X, t) \quad (26)$$

These rule are intuitively proven, and can be deduced from the definition of the temporal modalities. Using this working memory we can get the values of the current system variable, if no information of the system variable exists in the working memory the variable holds the same value as the last time point. The main algorithm is interactive and runs as the process of crisis management goes, so in each real time point the algorithm is executed to get the suggested decisions in this time point, the following steps represents the main process:

1. update the working memory to match the current time point
2. Get the current variable values, either from the working memory or from a measurement source.
3. Check the available resources.
4. Check the temporal events and determine the possible actions
5. For each action in the possible action list: Run the prediction algorithm assuming the action is taken.
6. Evaluate the results of the last step, suggest actions, answer queries, and get the user's decision.

It is clear that the system proposed is an advisory system [11], it doesn't suggest the final decision, However, it suggest actions, show clarifications and leave the last decision to the user in each time point.

The prediction algorithm is similar to the main algorithm, but it advances the time virtually, it goes in the future, and test

all possible paths, that is, in each decision point; it tests both situations when action is taken and when it isn't. the algorithm goes in a fixed extent in the future, when getting to the end of this extent it calculates the target of the system, and consider it as an output, the steps of prediction algorithms are:

1. If it is the last prediction time point return the system goal $Dis(\square C, T)$, else: 2. update the working memory to match the current time point.
3. Get the current variable values, either from the working memory.
4. Check the temporal events and determine the possible actions.
5. For each action in the possible action list: Run the prediction algorithm assuming the action is taken in the next time point.
6. Evaluate the results of the last step, choose best action, and return the system values accordingly. Apparently the algorithm searches all available future line, the complexity of the algorithm is $O(2^n)$ where n is the number of actions that may be taken in the future. To enhance the algorithm complexity we can use the concept of *classification*. For each current environment state vector there is a best action to be taken, where similar cases can have the same action, so we can say: *If* $Vsim(V_1, V_2) < \text{Threshold}$ then action taken in the case of V_1 can be taken in the case of V_2 .

$Vsim$ is the vector similarity function defined in 5, the value of Threshold can determine how reliable is the system, small value makes it more reliable but with high complexity, whereas low value means the opposite.

IX. THE GENERALIZATION OF THE MODEL

The crisis elements and environment variable models described so far, can suite a lot of types of crises, to generalize the model so that it can be applied to most crisis situation; we can use the concept of object, an object is an element of the crisis environment, each object has a variables and those variables are linked temporally, events are connected to objects, when an event is linked to a variable of an object, then it is connected to this object, the action can be applied to objects when events of these objects satisfy the action definition, resources are linked to actions and their consumption rules are related to the objects of the action. All of these information is fed to the system as meta model of the environment, then a model can be built upon this meta model, this meta model can suite most crisis environment, the meta model looks like the follows:

```
Object o_name (properties){ Variables,Events :
event_definition}
Resource name{Resource consumption rules}
Action{Events of objects related, Resources needed,Objects
variables change}
Global temporal relations{Left hand side events, Consequents}
```

X. CONCLUSION

In this paper we show how a logical model of a crisis can be built using the fuzzy representation and temporal

logic, characteristics of crisis environment has been modeled, a model similar to temporal linear logic has been used to represent resource availability and resource reusability, fuzzy inference techniques has been generalized to add temporal modalities, a prediction algorithm has been introduced and used to expect the future of the environment state, the idea is based on temporal simulation, the main benefit we get is that we don't need an expert to describe the temporal relations, they represent the real life relations, whereas the system uses them to simulate the flow of events and hence can provide decision support and resource management.

REFERENCES

- [1] Edward S. DELVEN, *Crisis Management Planning And Execution*, Auerbach Publications, 2007.
- [2] Gaberiel Prelipean, Mircea Boscoianu *Emerging Applications of Decision Support Systems (DSS) in Crisis Management*, Efficient Decision Support Systems - Practice and Challenges in Multidisciplinary Domains, ISBN 978-953-307-441-2, 2011.
- [3] Huda Nokhbatolfighahaayee, Mohammad Bagher Menhaj, Masoud Shafiee, *Fuzzy Decision Support System For Crisis Management With a New Structure For Decision Making*, Expert Systems With Applications 37 3545-3552, 2010.
- [4] Daniel J. Power, *How can DSS help in crisis planning, response, and management*, Power enterprises, 2012.
- [5] Ropert Babuska, *Fuzzy And Neural Control*, Deift University of Technology, The Netherlands, 2000.
- [6] Beg Ismat, Ashraf Samina, *Similarity Measures for Fuzzy sets*, Computer & Mathematics with applications International Journal, Vol. 8. 192-202, 2009.
- [7] M Fisher, D Gabbay, L Villa *Handbook of Temporal Reasoning in Artificial intelligence*, ELSEVIER, 2005.
- [8] Gerard Jean-Y, ves. Marselille *Linear logic: Its syntax and semantics*, Laboratoire de mathematiques discretees.
- [9] Pham Duc Q, Harland James, Winikoff Mickael, *Modeling Agents' Choices in Temporal Linear Logic*, Melbourne, Australlia : RMIT university, 2007.
- [10] Orchard Bob, *FuzzyCLIPS Version 6.10d User Guide*, Integrated Reasoning Group, Institute for Information Technology, National Research Council Canada, 2004.
- [11] Burstein, Frada and Holsapple, Clyde W, *Handbook on Descision support systems*, Springer, 2008.

Recognition of the Patterns Represented as the Field Structures

A.O. Chechel, andrey_chechel@mail.ru

The Bonch-Bruевич Saint-Petersburg State University of Telecommunications
61 Naberezhnaya reki Moyki
Saint-Petersburg, 191186 Russia

Abstract—The paper describes developed concept of field model based on Green's function. It allows converting the pattern edges into linear objects which are used particularly as a basis for the image geometrical mask construction. The introduced approach enables increasing of quality and performance for image processing tasks due to special pixel chains analysis and embedded mechanism for reduction of noise and image distortion. The concept is applicable to pattern recognition, image compression and video tracking processes.

I. INTRODUCTION

The problem of pattern recognition can be divided into the following tasks: the original image segmentation and classification of the selected segments. There are quite a number of different approaches of image segmentation: threshold techniques, edge construction, region growing technique including centroid binding, merging-splitting and watershed methods; texture techniques [1–2]. In this article we will review a developed segmentation algorithm based on the edge construction approach.

The main feature of this approach is to use the Green's function which allows us to extract a geometry mask from the image, thus making it possible to process it in the plane of figure. Visual information is represented as a scalar field. Such representation allows selecting those pixels from the image which have the exceeded predetermined threshold differences in the color gradations in their neighborhood. And interpret them as sources of coloration such as a dipole or a singular point (the power of these sources is proportional to the values of the differences of color gradations). Further application of Green's function allows constructing a field of singular points, which include all the differences of the color gradation greater than a predetermined threshold. In other words, the color-grade of the field constructed in such way is completely determined by differences of color gradations around the sources. It is obvious that subtractive field, obtained by subtraction of the geometry mask from the source image, does not contain any differences of the color gradations more than the given threshold. If we put away all

the nuances concerning the ratio of window size, in which is determined the impact on the total field of each source and the smoothing parameter [3], than we can say that if the predetermined threshold tends to zero, the subtractive field tends to smoothed (averaged) field.

II. CONCEPT OF THE FIELD MODEL

The essence of the approach is to divide the image in two different by type parts: the field of singular points, which is a mask of the image, and subtractive field, which is color base. These parts are processed in a fundamentally different ways. In the case of the geometry mask, a Green's function (the field of point source) is determined for the chosen equation of state. This gives us an opportunity to calculate the total field of all singular points, which is a particular solution of the chosen equation of state. For example, for the two-dimensional Poisson's equation we will have:

$$\frac{\partial^2 V}{\partial x^2} + \frac{\partial^2 V}{\partial y^2} = \sum_{j=0}^n \left(k_j^x dU_j^x \delta'(x - x_j) \delta(y - y_j) + k_j^y dU_j^y \delta(x - x_j) \delta'(y - y_j) \right). \quad (1)$$

In (1) V is image field; j is a summation index for the singular points; n is a number of special points; dU_j^x, dU_j^y are differences of the color gradations in x and y axis, k_j^x, k_j^y are the normal projection on x and y (for edges formed by pixels these parameters are set to either 0 or 1). As we can see from (1), in order to obtain a particular solution, we need to process all points of the image, which have differences in color gradation more than the given threshold. So in case of such representation the field of singular points is actually a geometry mask of the image. In the sense that while differing by color-grade from the source image, it includes substantial differences in color (note that the difference between the original image and the field of singular points is a smooth field). Using a special chain code [4], the singular points are grouped in a sequence of pixels (edges), combined from the

neighboring pixels by the criteria of the color gradation nearness.

This approach gives us a compact representation of the information about singular points. By adding to the chain code the coordinates of the first pixel, than starting from it, we will be able to restore the whole linear object. Further the implementing of the Green's function allows us to adequately convert linear objects to their original flat fragments. This makes the field model convenient for the pattern recognition, and also for the compact information storage. It should be noted that in the three-dimensional field model, edges are also a linear objects, but in this case they depend on the three coordinates. Thus, having performed the segmentation of the image using the concept of the field model, we can get a set of linear objects represented in the form of chain codes that uniquely describe the geometry mask of the source image. The described segmentation process is shown in the Fig. 1.

In order to implement the analysis of the single linear segment there was developed an algorithm for finding the angle in the chain of pixels. The main feature of the algorithm is the adequate processing of the objects that are located at different angles. The main point is that straight lines on the images are displayed with distortions, except for strictly horizontal and vertical (if we consider a single pixel and its neighborhood). This introduces a number of nuances in the process of detection angles in the chain of pixels. We must consider all possible distortions while detecting angles. For this purpose we applied a method of finding an angle using a window. When the deviation of a straight line within the window exceeds a predetermined threshold, then we can say that here lies an angle. The number and location of the angles in the chain code is a rough description that can be applied to almost all graphic objects. And it allows us to perform initial filtration of the analyzed objects.

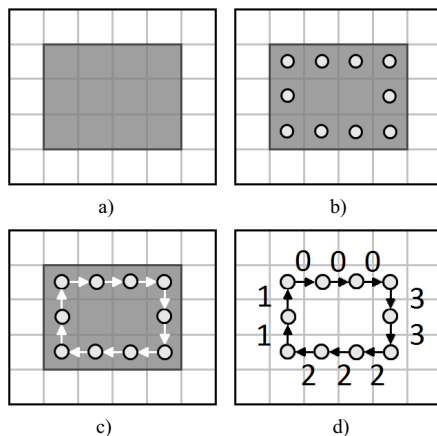


Fig. 1. The process of forming a linear object: a – source image; b – a selection of singular points; c – merging the singular points in a cyclic chain code; d – referring the digital values to the chain items.

So as a result of the segmentation algorithm the scene is decomposed into a number of geometrical linear objects with the set of properties inherent to each single object. Further analysis of the scene is to analyze this set of objects. Each linear object can be both a whole real object and a part of it. Furthermore, there may be cases when one object overlaps the other. These facts can make the analysis significantly more difficult. However, to avoid the problems described above the obtained primary objects undergo the processes of merging/splitting. Primary objects are united in a group, or conversely, decomposed into more simple objects adjoined to the other group of objects. Each group is a template of a real object, which consists of primary objects, and can identify the real object that it represents. At this point, objects should be filtered in order to select only those that satisfy the conditions of the task.

III. CHAIN ANALYSIS

One of the basic algorithms for identifying patterns by analyzing the edges of an object is a method of comparing the two chain codes with the definition of the coefficient of similarity of the geometric shapes of the objects. Since there are always distortion and noise on the images, than it will be not enough just to track and compare all items of the chain code in order to solve this problem. In practice, it is more convenient to move from the work with the elements of the edges to analyses of the linear graph. The distance from the center of the object to each particular point of the chain is taken as function value. The resulting function is continuous, moreover, it can be considered as a periodic, taking into account the fact that the contour of the object is closed. The function allows comparing not only the particular items of the chain, but also local changes in the neighborhood of the concerned element. This approach is superior to sequential comparison of the chain items, as distortion of the object, if any, usually affects a group of items at once and should be considered together in the analysis of the object. If we use additional parameters like statistical expectation of the random value and turning moment about axis, it will allow us to take advantage of linear objects. While processing one-dimensional arrays of data we can ignore minor distortions and inaccuracies and respond only to substantial deviations. As an example we suggest using this algorithm in the classification of the symbolic information. While the existing methods of image processing allow us to work with objects that are identical to the standard samples, the methods developed in the field model, allow a partial compliance with the sample, which can be expressed using a different font, italic, etc.



Fig. 2. Car identification within the video tracking system based on the field model concept.

Methods of contour analysis are very demanding for the performance of hardware. However these methods are much more efficient. This is because the analysis of the linear objects is speeded up by the wavelet analysis applied to linear objects. This greatly reduces the overall amount of information and allows us to focus only on the important characteristics of a linear object. This factor makes it possible to use these methods in video processing as well. The system of video tracking described in [5] is based on the concept of field models for frame-by-frame processing and also for analysis of the interframe information (obtained by subtracting subsequent frame from the previous one). The fact that this concept allows to uniquely present an image geometry mask in the form of set of linear objects, makes it possible to use it both for analysis of the primary frame, and of the subtractive interframe information. The use of the obtained linear objects enables us to reduce by several times

the amount of the data being processed. Hence the time spent on the entire process decreases too. Fig. 2 shows an example of the work of the video tracking system. A moving car in the center of the frame is set as a target object. The system identifies the object, as evidenced by a rectangular frame around the car.

IV. CONCLUSION

The proposed model can be used in a wide class of images. The mechanism of association of the singular points in the chain is based on the proximity of the color gradations of neighboring points. The multitude of the constructed edges forms a geometry mask of the image that can be used apart from the subtractive field. The fact that the edge is represented by a linear object describes a flat graphic object allows us to carry out such processing as: pattern recognition, morphological analysis, video tracking.

REFERENCES

- [1] E.P. Putyatin and S.I. Averin, "Image processing in robotics," Moscow: Mashinostroenie, 1990.
- [2] R.C. Gonzalez and R.E. Woods, "Digital Image Processing," 2nd ed., Upper Saddle River NJ: Prentice Hall, 2005.
- [3] J.F. Boltov and I.A. Volkov, "Image smoothing based on solving of boundary value problems," Telecommunications, vol.5, Moscow: Science and Technologies, 2010, pp.24..33.
- [4] J.F. Boltov, "Compression of graphical information on basis of its presentation in the form of field structure," Telecommunications, vol.12, Moscow: Science and Technologies, 2008, pp.30..35.
- [5] A.O. Chechel, "Graphic object tracking by representing as the field pattern," 20th International Scientific and Technical Conference "Modern Television and Radioelectronics", Moscow: FSUE MDB Electron, 2012, pp.130-133.

Analyzing Reflector Effect on SAR Imaging System with a Proposed Functional Model

Mojtaba Behzad Fallahpour
 Young Researcher Club of Lahijan Islamic Azad
 University, Iran
 m_behzad_fp@yahoo.com

Hamid Deghani
 Islamic Azad University of Buser, Iran
 hamid_deh@yahoo.com

Abstract— Synthetic Aperture Radar (SAR) is active and coherent microwave imaging system which has the capability to image in various climatic conditions and darkness. Due to special usage of SAR system in military, opposing with it through passive defense mechanisms seems to be necessary. One effective way to oppose with this powerful system is reflector. Reflectors have wide radar cross section (RCS) in comparison with their sizes, so amount of waves' reflection from surface of things and phenomenon in region in comparison with amount of reflectors' reflection is negligible. This paper, at first will try to extract two-dimensional impulse response from SAR imaging system to simulate receiving raw signal from desired region, and then determine effectiveness of reflector in region on SAR images. The simulation results will prove the effectiveness of reflector to oppose with SAR.

Keywords- SAR; Impulse Response; Raw Data; Reflector;

I. INTRODUCTION

Remote sensing is science of acquiring and interpreting of information from different targets without direct contact with them. This science is used frequently in many different scientific and researching fields; like geology, mine, cartography and etc. Generally, Remote sensing systems are divided into two groups; active and passive [1]. SAR is an active Remote sensing system. High resolution images of the Earth covering large areas can be obtained using satellite or airborne SAR [2]. Some abilities of this kind of measurement systems are:

- Ability of taking image in darkness
- Ability of taking image in various climatic conditions
- Ability of recognizing slope
- Ability of imaging with various polarizations
- Ability of recording of sending and receiving waves' phase and as a result ability of detection done activities under surface

So all countries are studying and researching to access this important technology, because beside mentioned usages, SAR has important usage in military; by accessing this technique, any country can be able to take high quality images from crucial areas of other countries, so from a passive defense standpoint opposing with this system is significant and vital [3].

This paper is organized as follows: in section II, a brief back ground of reflector is presented. Extracting of impulse response of SAR imaging system is described in section III. Finally in section IV simulation of reflector's effect in opposing with SAR imaging system will be presented.

II. REFLECTOR

The reflectors are considered as one of most important opposing ways to radar imaging systems; because this kind of fake targets have wide radar cross section (RCS) in comparison with their sizes. These systems have simple structures and usually are made in the shape of pyramid with metallic plate (for example aluminum).

If these targets be used in region, the amount of wave reflections from surface of things and phenomenon in region in comparison with amount of reflection of reflectors is negligible and SAR image will be made by wave reflections from surface of fake targets (reflectors) not things and phenomenon in region. The general structure of reflectors is shown in figure 1. As can be seen in this figure, reflectors can be made in different shape and be used. Any shape has its special amount of reflection; and amount of reflection for different wave length can be determined and also can be decided which kind of fake target is more suitable and effective for each wave length to hide targets.

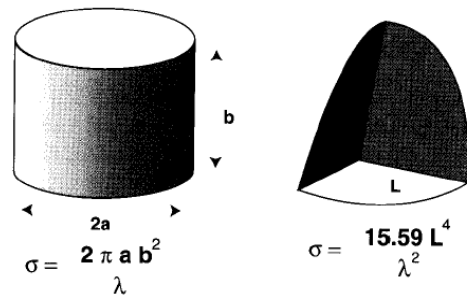


Figure 1. The samples of reflectors (σ is radar cross section of reflectors) [5].

To oppose with SAR imaging systems, a number of reflectors, with proper distribution, can be placed in desired region. Each SAR imaging system, that images from this region, will take an image with high brightness which does not show any things or phenomenon [4].

III. IMPULSE RESPONSE

Figure 2 shows a simple geometric model of the SAR location and the beam footprint on the Earth's surface.

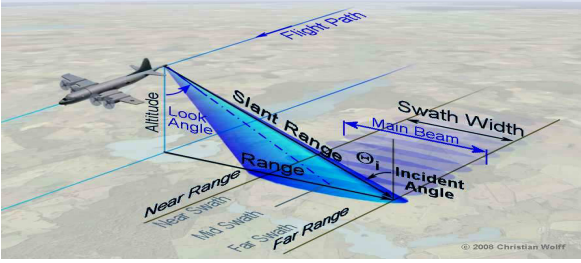


Figure 2. SAR geometry [6]

In SAR system each point on the ground, between near range and far range, is illuminated by the beam for a duration of T_r . The reflected energy at any illumination instant is a convolution of the pulse waveform and the ground reflectivity, g_r , within the illuminated patch as:

$$s_t(t) = g_r(t) * s_{pul}(t) \quad (1)$$

This energy arrives back at the receiving antenna between times $2t_1$ (near range time) and $2t_2$ (far range time) thereby range.

Consider a point target at a distance, R_a , away from the radar with a magnitude, A_0 which models the backscatter coefficient, σ_0 . This means that:

$$g_r(t) = A_0 \delta(t - \frac{2R_a}{c}) \quad (2)$$

Where $2R_a/c$ is the delay time of the signal for that reflector. The demodulated baseband signal from a single point target can be presented by the complex signal:

$$s_0(\tau, \eta) = A_0 w_r \left(\tau - 2 \frac{R(\eta)}{c} \right) w_a(\eta - \eta_c) \times \exp \left\{ -j4\pi f_0 \frac{R(\eta)}{c} \right\} \times \exp \left\{ j\pi k_r \left(\tau - 2 \frac{R(\eta)}{c} \right)^2 \right\} \quad (3)$$

Where the coefficient A_0 is a complex constant and w_r is pulse envelope in range direction. The pulse envelope is usually approximate by a rectangular function as:

$$w_r = \text{rect} \left(\frac{\tau}{T_r} \right) \quad (4)$$

Where T_r is the pulse duration and τ is the fast time (time in slant range direction). w_a is beam pattern of antenna in azimuth direction and η is slow time (time in azimuth direction). k_r is the FM rate of the range pulse. If A_0 is ignored, (4) is the impulse response of a point target having unity amplitude. Thus, the important SAR sensor

impulse response is given by

$$\text{himp}(\tau, \eta) = w_r \left(\tau - 2 \frac{R(\eta)}{c} \right) w_a(\eta - \eta_c) \times \exp \left\{ -j4\pi f_0 \frac{R(\eta)}{c} \right\} \times \exp \left\{ j\pi k_r \left(\tau - 2 \frac{R(\eta)}{c} \right)^2 \right\} \quad (5)$$

To model the signal received from a general ground surface, the ground reflectivity is convolved with this impulse response in two dimensions to give the baseband SAR signal data as:

$$S_{bb}(\tau, \eta) = g(\tau, \eta) * \text{himp}(\tau, \eta) + n(\tau, \eta) \quad (6)$$

Where $n(\tau, \eta)$ is an additional noise component that is present in all practical systems. The noise originates mainly from the front end receiver electronics, and can be modeled as Gaussian white noise. The SAR system model corresponding to (6) is shown in Figure 3[7, 8].

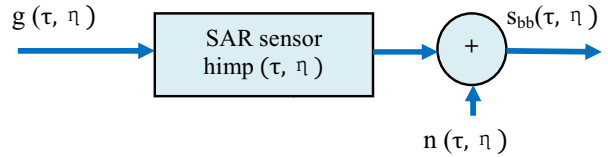


Figure 3. The SAR system model [7]

In this paper, to extract the impulse response of SAR system, a point target is determined into empty image (a target with a size of 1 pixel of SAR image) and to simplify, this target is considered to make maximum light for SAR image and also be seen equally in all directions. In the other word, impulse function is made in two dimensions and this impulse image is shown in figure 4-a.

Usually three kinds of algorithm, in SAR imaging system, are used to transform radar raw data to image. These three image formation algorithms are RAD, CSA, and OMEGA-K. In this paper, CSA (Chirp Scaling Algorithm) algorithm is used. This algorithm is performed on impulse image conversely, and image of impulse response of SAR imaging system is extracted [9, 10]. RADARSAT-1 characteristics are used to extract impulse response of imaging system. This impulse response image in two direction, range and Azimuth is shown in figure 4-b.

Now the passive defense actions can be simulated by using of impulse response. In fact this impulse response is a CAD (Computer Aided Design) for transforming image to data. For example, optical image from a region will be convoluted by impulse response and radar raw data of region will be extracted. Then, radar data will be transform to image with imaging algorithm and radar image of region will be obtained.

To show accuracy of obtained impulse response, CSA algorithm is used to convert impulse response to impulse function and it is shown in figure 4-c. This figure shows the accuracy of done performance.

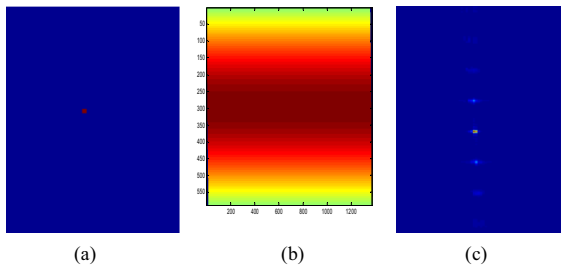


Figure 4. a) Generating of impulse function. b) Impulse Response of SAR. c) Image of Impulse response with CSA.

IV. SIMULATION AND RESULT

To simulate effect of reflector in opposing with SAR imaging system, the image is chosen from north of Iran (Guilan-Astane Ashrafiye) and reflectors are used in two positions. In first position, entrance of large square is camouflaged by reflectors. In second position, first, the building of hospital is camouflaged and then a few fake positions are created in vicinity.

Camouflaging of entrance square by reflectors

Figure 5-a, b are respectively optical and radar image of Astan e Ashrafiye in Guilan. The entrance square is shown in red box. The purpose is camouflaging square by reflectors. By using of “MATLAB” software, the square is covered and simulated in SAR image.

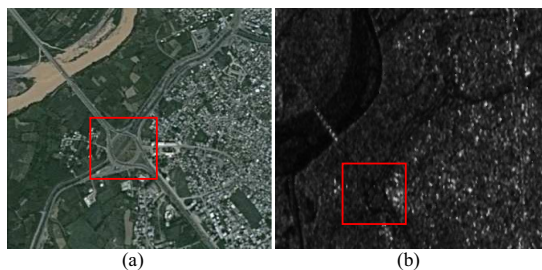


Figure 5. Optical image, b) Radar image from north of Iran (Guilan-Astane Ashrafiye)

The original image (figure 5-b) and combined image with reflector are transformed to radar raw data by using of obtained impulse response in previous section. Amplitudes and phases of these data are shown in figure 9. Differences between amplitude and phase by original image and image with fake targets are visible. Now to prove accuracy of performance, all above stages will be done conversely by CSA. As shown in figure 6-a, b, restored images have proved using of reflectors to oppose with SAR imaging system. For clarity, Comparison of restored images is presented in

figure6-c, d. In figure 6-c, the square is obviously clear, but in 6-d the square is vanished under reflectors.

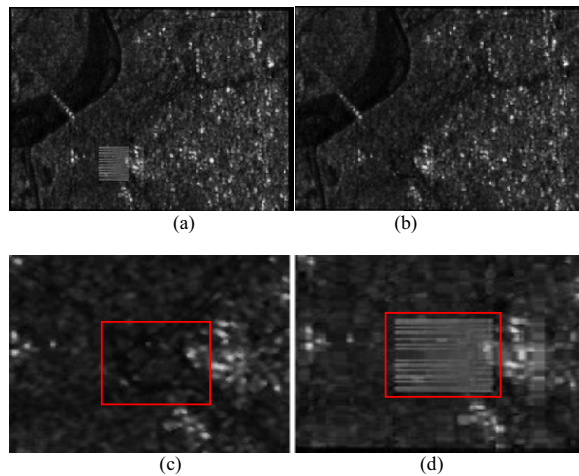


Figure 6. (a) Restored image from raw data of combined image with reflectors, (b) Restored image from raw data of original data.c,d)Comparison of restored images , before using of reflector and after using of reflector

Creating few fake positions

Consider figures 7-a, b. The Kosar hospital in Astane Ashrafiye is shown by red box. The purpose is camouflaging of hospital (box 2) and creating of fake targets (box 1) by using of reflectors. In figure 8, points of fakes are created in box 1 and also camouflage is created in box 2 using of “MATLAB” software.



Figure 7. a)optical image of Kosar hospital in AstaneAshrafiye. b) Red box 1 is place of creating of fake targets, and red box 2 is place of creating of camouflage

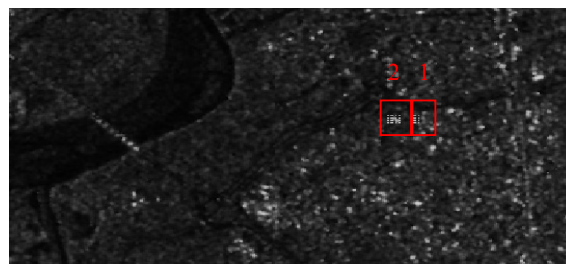


Figure 8. Creating of bright points in image by reflectors to make fake points, in box 1 and camouflage in box 2

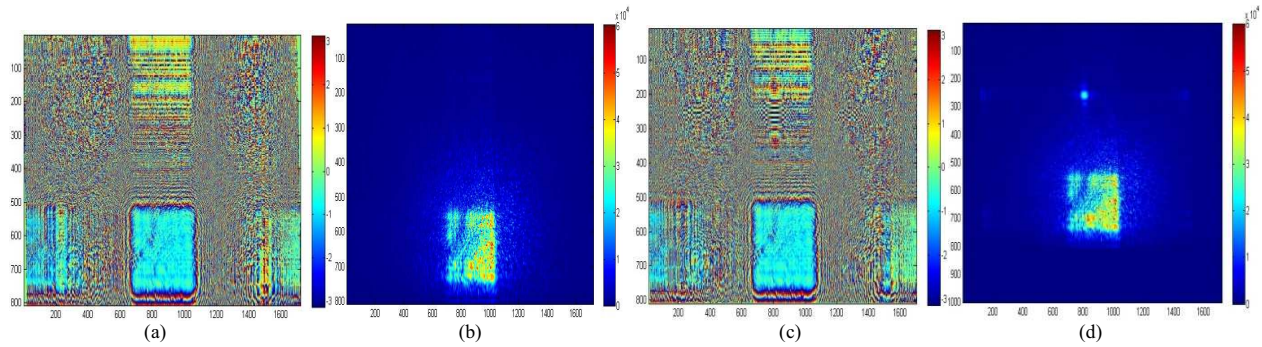


Figure 9. (a) Phase of raw radar data of original image, (b) amplitude of raw radar data of original image, (c) Phase of raw radar data of combined image with reflectors, (d) amplitude of raw radar data of combined image with reflectors.

The combined image by reflector (figure 8) is transformed to radar raw data by using of obtained impulse response in previous stage. After changing of above images to radar raw data, the amplitude and phase of this data is shown in figure 10.

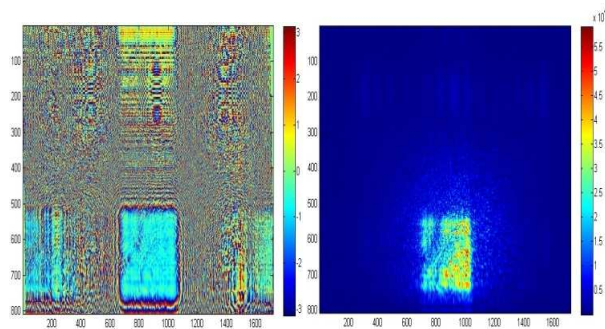


Figure 10. (a) Amplitude, (b) phase of raw radar data from combined image with reflectors

Now to prove accuracy of operation, the above operation will be done conversely by CSA. As it shown in figure 11, the restored image proves accuracy of operation, obtained impulse response and using of reflector to opposing with SAR imaging system.

For clarity, Comparison of restored images is presented in figure 12. In figure 12-a, the hospital is obviously clear and in box 1 situation is normal. But in figure 12-b, the hospital is fully camouflaged and in box 1 creating of fake structure is shown.

CONCLUSION

This paper, by using of two-dimension impulse response of SAR radar, is shown that one of the most effective mechanisms to oppose with powerful SAR imaging system is using of reflector. Because reflectors create wide radar cross section than their sizes and also, by using them, amount of wave reflection from surface of other things and phenomenon in comparison with amount of fake targets' reflection is negligible.

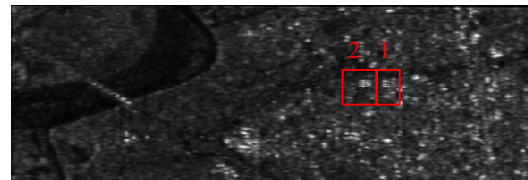


Figure 11. Restored image from raw data after using reflectors

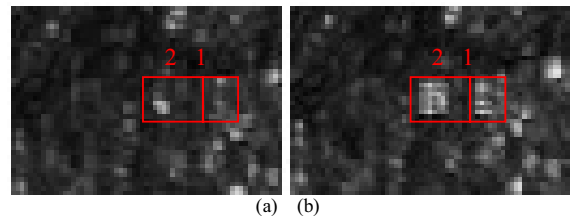


Figure 12. Comparison of restored images, (a) before using of reflector and fake, (b) after using of reflector and fake

REFERENCES

- [1] J.A.Richards, Remote Sensing with Imaging Radar, 1st edition, Springer, New York, 2009.
- [2] S.Bhattacharya, T.Blumensath, B. Mulgrew, M. Davies, "FAST ENCODING OF SYNTHETIC APERTURE RADAR RAW DATA USING COMPRESSED SENSING", in IEEE conference, pp.448-452, 2007.
- [3] J. Stimson, Introduction to airborne radar, 2nd edition, SciTech California, 1998.
- [4] J. Zhao, C.Li, J.Yin, X.Shan, G.Zhang, F.Ye, "Effects Analysis of Machining Tolerance on RCS of Corner Reflectors", International Journal of Digital Content Technology and , pp.91-97, July 2011.
- [5] A. W. Doerry, "Reflectors for SAR Performance Testing", Sandia National Laboratories reports, January 2008.
- [6] http://www.radartutorial.eu/20.airborne/pic/SLAR-geometry_p.jpg.
- [7] Cumming, I.G.; Wong, F.H. "Digital Processing of Synthetic Aperture Radar Data Algorithms and Implementation." Artech House: London, 2005.
- [8] H. Maitre, Processing of Synthetic Aperture Radar Images, 4th edition, Wiley, New York, 2011.
- [9] A.S. Khwaja, L. Ferro-Famil, E. Pottier, "SAR Raw Data Generation Using Inverse SAR Image Formation Algorithms", in IEEE conference, pp.4174-4177, 2006.
- [10] Z.Donghoa and Z.Xiaoling, Downward-Looking 3-D linear Array SAR Imaging Based on Chirp Scaling Algorithm, in IEEE conference, pp.1043-1046, 2009.

Educational and pedagogical design of a software tool for learning postero-anterior cephalometric landmarking.

Francesco Maiorana¹, Rosalia Leonardi²

Department of Electrical, Electronics and Computer Engineering¹, Department of Orthodontics²
University of Catania, Viale Andrea Doria, 6 – 95125 Catania, Italy
francesco.maiorana@dieei.unict.it¹, rleonard@unict.it²

Abstract- One of the key instruments for orthodontic patient evaluation is the cephalometric analysis that is used for diagnosis, treatment planning and evaluation. In clinical practice it is important to train students for a correct cephalometric analysis in order to master both lateral and postero-anterior cephalograms. The combination of lateral and postero-anterior analysis for a multimodal analysis is gaining importance in order to be able to evaluate the anatomical structure also in the frontal plane which allows for a better evaluation of asymmetry and craniofacial widths. In this work we present the educational and pedagogical idea that inspired the design of the software tool for learning postero-anterior cephalometric landmarking. The tool can be used in conjunction with a tool for latero-lateral cephalometric analysis offering a blending of approaches that can offer, in a multimodal way, a broader view and evaluation of the clinical situation.

I. INTRODUCTION

Cephalometric radiography was introduced in research and clinical orthodontics because with skull radiographs taken under standardized conditions, the spatial orientation of different anatomical structures inside the skull can be studied more thoroughly by means of linear and angular measurements [1]. Cephalometric analysis is used for treatment planning and evaluation. The major sources of error in cephalometric analysis include systematic errors such as radiographic film magnification, and random errors such as variation in tracing, measuring, recording and landmark identification [2]. The inconsistency of landmark identification depends on the experience and training of the clinician. Performing manually a cephalometric analysis is a very time-consuming process. In recent years various commercial software has been developed to perform computer-aided cephalometric analysis, in which the operator digitizes the landmarks and then the software proceeds to draw the lines and to compute the needed measurements. This eliminates the errors due to line drawing and to measurements with a protractor. The computerized analysis can be completed significantly faster than a manual one. It also potentially increases the accuracy of the obtained result, as demonstrated by comparison between manual tracing and direct digitization, but this demands radiographs of high quality and an operator with detailed knowledge of radiographic anatomy [3]. Accuracy is a crucial issue since measurements are considered precise when errors are within 1 mm (the mean estimating error of expert landmarking identification has been reported to be 1.26 mm [4]); measurements with errors within 2 mm are considered

acceptable, and are used as a reference to evaluate the recognition success rate of an automatic software, although the former level of precision is deemed desirable [5].

The software for computerized cephalometric analysis are still prone to errors in landmark identification. Moreover, both manual and computer-aided cephalometric analysis are, at a different level, time consuming.

The request for automatic or semi- automatic landmark identification has become urgent with the goal of reducing the time required for landmark identification and to assist the experts by proposing a set of identified landmarks that eventually can be modified by the expert. This could lead to reduced time for the complete analysis with a reduction of the errors an improvement of accuracy [6]. The tool involving automatic landmark identification could be a useful tool in assisting clinicians in everyday practice.

The problem of automatic landmark identification is a challenging one [7-10] and involves expertise in different fields ranging from artificial intelligence to computer vision.

In clinical practice it is important to train students for a correct cephalometric analysis that include mastering both a lateral and a postero-anterior cephalogram (PAC). For a tool successfully used in learning lateral cephalogram the reader can reference [11-13].

The importance of PAC in clinical practice has recently been reviewed in [14] and the accuracy in landmark identification on digital PAC has been assessed in [15].

This work presents an educational learning tool that can be used to practice and acquire a deep understanding of PAC landmarking. The key pedagogical aspects of the platform that have guided its design will be presented as well as its main capabilities and a first evaluation in a field study.

This work is organized as follows: section 2 contextualizes the problem and briefly reviews other educational tools, section 3 describes the educational and pedagogical design of the software tool, section 4 presents the software tool and its evaluation, both qualitative and quantitative and section 5 draws some conclusions and highlights future work.

II. PROBLEM CONTEXTUALIZATION AND STATE OF THE ART

One of the key aspects in orthodontics is malocclusion diagnosis and treatment. For the correct management of the malocclusion problem the anatomical structures of the patient have to be evaluated in all the three planes and precise landmark localization has to be performed in order to take exact linear and angular measurements.

In a multimodal analysis [16] the combination of lateral and postero-anterior analysis is gaining importance in order to evaluate anatomical structure also in the frontal plane in order to allow for a better evaluation of asymmetry and craniofacial widths [17]. The blending of approaches of different types of analysis, can offer, in a multimodal way, a broader view and evaluation of the clinical situation.

For these reasons the educational process must address learning and practicing in the landmarking process in both lateral and postero-anterior orientation and the educational approach should reflect modern teaching guidelines allowing for active, self-regulated learning in order to acquire a deeper knowledge linked to clinical practice and diagnosis.

The educational approach should also reflect the necessity to creating effective and engaging information literacy curricula in dental higher education [18] where the key aspect is not the ability to access information but the ability to generate knowledge from the retrieved information. Students have to be able to analyze and synthesize information and analysis in a multimodal fashion in order to converge to an optimal diagnosis and treatment planning.

In this scenario it is essential to provide a software tool that through an interactive session offers to the students the possibility to acquire deeper and wider knowledge. An example in this direction is the work presented in [19] where the author presents a tool that is “a power point presentation consisting of numerous slides of varying transparency of hard and soft tissue anatomy superimposed on the same radiographic image. The interactive function built into the digital tool was the ability to scroll through these transparencies”. The tool also uses text and labels indicating the anatomical features described in the radiograph. The author reports a slight improvement in the average score post intervention in the students that have used the software tool compared with the student that learned through textbooks but in qualitative tool evaluation the author reports that a remarkable 94% of the 88 students preferred using the digital tool to the textbook.

A similar result was obtained in [13] where the author reports an increase, after tool usage, in performance measured in terms of precision measured as distance from user landmark and reference landmark, and in term of efficiency measured in terms of time needed to complete the analysis. The qualitative user evaluation of the tool showed a very high appreciation by 80 % of the six experts. The tool discussed in [13] must be viewed as a complementary tool that expands, in a multimodal approach, the tool presented in this work.

III. EDUCATIONAL AND PEDAGOGICAL DESIGN OF THE SOFTWARE TOOL

The software tool was designed with these educational goals in mind:

- Allow for ‘learning by doing’ and ‘active learning’ engaging students in practical exercises linked to relevant clinical activity
- Offer the possibility to students to learn at their own pace, thus avoiding cognitive overload considered as one of the main reasons of failure and making

possible self-directed practical learning focused on interactive and quality practice and feedback .

- Offer the possibility of interactive landmarking sessions allowing for in-depth reflection, repetition and comparison with the expert landmark.
- Allow for self-evaluation and assessment of progress both in terms of precision, measured as distance from an expert landmark and in term of efficiency measured as time needed to perform the landmarking time and the number of re-do operations. The student performance can be profiled and compared also with objective peer performance, a feature that is appreciated in learning .
- Give immediate and objective feedback automatically generated by comparison with the gold standard landmarking of the x-ray resulting in time-saving for the professor without loses in quality.
- Offer to the student a database of clinically relevant cases also covering complete patient history from diagnosis to intermediate treatment results and treatment completion and also a database of child cephalometries that are recognized as more difficult cases. The clinical case are arranged at different levels of difficulties both in terms of clinical conditions, patient age and image quality.
- Offer the possibility to visualize, as an animation, the sequence of mouse movements performed by the expert in the landmarking process, thus giving an indication of the search strategy and allowing for a richer practice focused on quality not quantity.
- Offer to the students image enhancement algorithms that can be used by a novice as an aid for their first attempt in the landmarking process and facilitating their task by enhancing the most difficult anatomical region.

The most suited pedagogical setting for the tool is a blended learning approach where the starting point for the learning experience is the classroom setting and then the use of the learning tool for active, self-directed practical and clinical relevant case training followed by recap sessions to give appropriate feedback and highlighting the most important concepts as well as the most common errors and how to avoid them and the most correct landmarking processes.

IV. THE LEARNING TOOL

An in house tool was written in Borland C++ version 6 to shows the x-ray and save the landmark coordinates with the possibility to make changes in the landmarking process. The expert located the landmark on the screen by a mouse click on the selected location. A pointing arrow cursor was used. The selected landmark is displayed by a red dot in the interface. The sequence of points to be landmarked was proposed in the software interface by the software based on a consensus sequence between all the experts and the medical coordinator of the project. The tool allows one to practice with thirty-four commonly used PAC landmarks. For their definition the user can reference [15].

The interface, as displayed in figure 1 distinguishes between left and right landmarks arranging the landmarks in

different panels thus avoiding confusion and difficulties in data interpretation. A confirmation message appears after the landmark is displayed on the screen allowing to the user to double check the landmark localization.

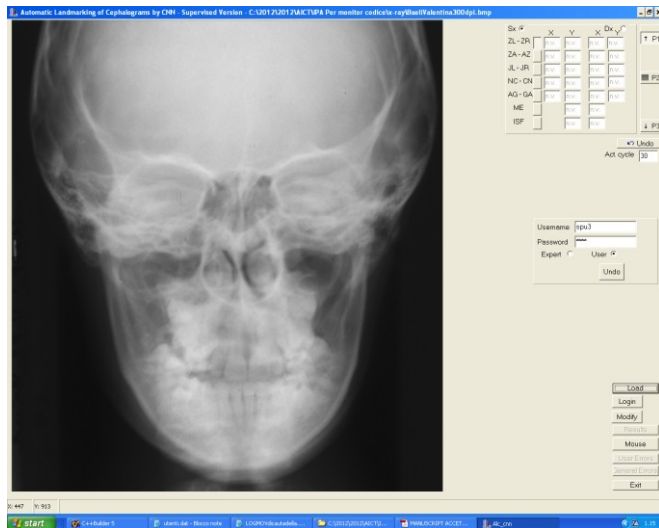


Figure 1: The educational tool interface

At the end of the landmarking session an overview of all the landmarks allows the user, with the aid of the software, to review and eventually correct the location of some landmarks.

An immediate feedback is provided by the platform that displays, on the same x-ray, the location of the landmark located by an expert and gives a clear indication of the amount of errors measured as distance between each user landmark and the expert one, distance measured both in pixels and in millimeters thus allowing for a clinical evaluation of the result.

Figure 2 shows an example of such immediate feedback where the yellow dots are the expert landmarks and the red dots are the user landmarks.

This type of feedback allows the user to critically reflect on his work and to objectively evaluate its performance both in terms of absolute errors and in terms of progression from previous training sessions.

The importance of feedback has received greater attention in academic literature. Recently in [20] the author suggests and highlights the importance of feedback that influences self-perception when it is considered by the students who consider it accurate, when they took responsibility for it and when it is motivating for evaluative judgments.

The tool supports a Cellular Neural Network (CNN) simulator with enhancing image algorithms that can also be used by a novice to enhance image quality or highlight difficult anatomical regions with, for example, low image contrast. The reader can reference [8, 12] for more details on the algorithms.

The tool also offers the possibility to visualize, as an animation, the sequence of mouse movements that are performed by experts while landmarking. The sequence of mouse movements is superimposed on the x-ray. An example of such movements is reported in figure 3.



Figure 2: Automatic feedback provided by the learning tool.

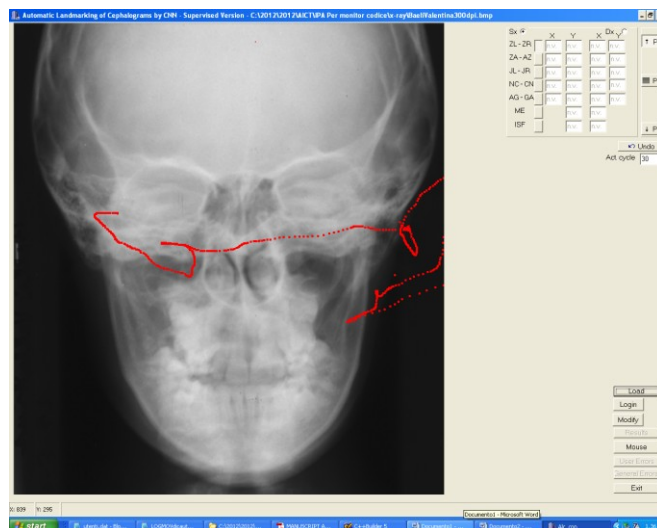


Figure 3: An example of sequence of mouse movement performed in the landmarking process.

A range of statistical analyses are offered to the users such as box plots of the error in each landmark over all the analyses, time charts of the errors in subsequent analysis, dispersion plots of the errors and Bland-Altman plots, Figure 4 shows an example of a Bland-Altman plot.

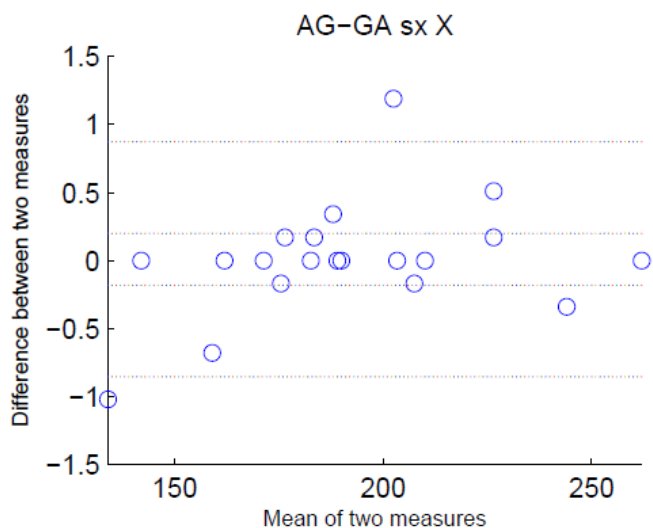


Figure 4: An example of the Bland-Altman plot for the abscissa of the AG-GA landmark

A Bland-Altman plot is used to verify the congruity and agreement of the x and y coordinate of each landmark. Statistical metrics can be presented to the user in order to guide him towards a better knowledge of the problem. The tool logs the time spent for each analysis, the time spent for each landmark, all the mouse movements and the number of redo operations for each landmark. All these measures can be used by the students to self-evaluate their performance through objective measures.

The tool was evaluated by an expert that landmarked all the points and her measures were used as reference landmarks and by other four users from the Orthodontic Department of Catania, Italy which were considered to have an equal level of knowledge of cephalometric tracing. The observers were all familiar with the software. Each observer landmarked all the 34 points in 20 cephalograms. One month after the experiment was repeated and the four students landmarked the same x-rays displayed in order to quantify retention and improvement.

The distance from the expert landmark and the student landmark is automatically evaluated and presented to the student both in terms of a single coordinate and as distance between the landmarks.

The tool was qualitatively evaluated by post interview with the expert and the students, and by a questionnaire. The result was encouraging since the tool was considered engaging, simple and easy to use.

To quantitatively evaluate the usefulness of the learning experience we computed the difference, measured in pixels, between the error in the first x-ray and the error in the last x-ray for each landmark. The results are reported in table 1.

From the table it is possible to appreciate the percentage of times each of the four students improved, where the improvement is measured as the decrease in errors computed in number of pixels. The percentage of improvements is always greater than 50% arriving at to 84% in one case.

The size of the differences between the errors in the first x-ray and in the last are comparable even if improvements are slightly greater.

TABLE I
STUDENT IMPROVEMENTS BETWEEN THE FIRST AND LAST X-RAY

User	% Number of improvements	Average improvement (pixel)	Average lose of precision (pixel)
U1	0,65	4,74	4,75
U2	0,68	4,30	4,58
U3	0,84	5,04	3,33
U4	0,57	5,19	4,38

V. CONCLUSIONS AND FUTURE WORK

In this work we have presented a learning tool that can be used to active student engagement in practical clinical relevant landmarking sessions that offer, in a blended learning approach, the possibility to practice in PAC landmarking with a high quality dataset and process.

As further study we plan a deeper evaluation of the tool, better integration with other tools for cephalometric landmarking such as [13], integration with tutorials explaining the anatomical structures and video-clips showing and commenting on the landmarking process and its relation with the undelaying anatomical structures.

Extension to a 3D learning environment and cephalometric landmarking could also be fruitful and can assist students in practicing and learning 3D spatial ability and interpretation of 3D images.

In the light of fostering the ability to generate knowledge from gathered information tools such as the one proposed in [21] can be used as an aid to cluster literature documents and discovering new associations. These clustered documents and the discovered associations as well as biomedical ontologies, such as disease ontologies, can be used to link the x-ray to the associated diseases or to the treatments.

REFERENCES

- [1] B.H. Broadbent "A new x-ray technique and its application to orthodontia", *Angle Orthod*, vol 51, 93-114, 1981
- [2] H.S. Kenneth, S. Tsang, M. S. Cooke, "Comparison of cephalometric analysis using a non-radiographic sonic digitizer", *European Journal of Orthodontics*, Vol 21 1-13, 1999
- [3] Yi.J. Chen, S.K.Chen ,H.F.Chang, K.C. Chen, "Comparison of landmark identification in traditional versus computer-aided digital Cephalometry", *Angle orthod*. Vol 70, 387-392, 2000.
- [4] S. Baumrind, R.C., Frantz, "The reliability of head film measurements landmark identification", *American Journal Orthod*, Vol 60, NO 2, 111-127
- [5] T. Rakosi, "An atlas and manual of cephalometric radiography", Wolfe Medical Publications, London, 1982.
- [6] T.J. Hutton, S. Cunningham, P. Hammond, "An evaluation of active shape models for the automatic identification of cephalometric landmarks", *European Journal of Orthodontics*, Vol 22, 499-508, 2000
- [7] R. Leoranrdi, D. Giordano, F. Maiorana, C. Spampinato, "Automatic cephalometric Analysis: A Systematic Review", *Angle Ortodontist*, Vol 78, No 1, 2008
- [8] R. Leonardi, D. Giordano, F. Maiorana, "An evaluation of cellular neural networks for the automatic identification of cephalometric landmarks on digital images", *Journal of Biomedicine and Biotechnology*, vol. 2009, 2009.
- [9] D. Giordano, R. Leonardi, R. F. Maiorana, C. Spampinato, "Cellular Neural Networks and Dynamic Enhancement for Cephalometric Landmarks Detection," *Proc. ICAISC 2006 Int. Conf. on Artificial Intelligence and Soft-computing*. Zakopane, Poland. Springer-Verlag, LNAI, vol. 4029, pp. 768-777, 2006.

- [10] D. Giordano, R. Leonardi, F. Maiorana, G. Cristaldi, M.L. Distefano, "Automatic landmarking of cephalograms by cellular neural networks", LNCS 3581, pp. 333-342, 2005.
- [11] D. Giordano, R. Leonardi, F. Maiorana, "Effects of Monitor Size on Accuracy and Time Needed to Detect Cephalometric Radiographs Landmarks," *Displays*, in press, doi: 10.1016/j.displa.2012.06.004 available on line at: <http://www.sciencedirect.com/science/article/pii/S0141938212000510?v=s>
- [12] R. Leonardi, D. Giordano, F. Maiorana, M. Greco, "Accuracy of Cephalometric Landmarks on Monitor-displayed Radiographs with and without Image Emboss Enhancement," *European Journal of Orthodontics* vol. 32, no. 3, pp. 242-47, 2010.
- [13] F. Maiorana, R. Leonardi, "A Cephalometric educational tool with expert feedback," *Proceedings of ITME*
- [14] R. Leonardi, A. Annunziata, M. Caltabiano, "Landmark Identification Error in Posteroanterior Cephalometric Radiography A Systematic Review," *Angle Orthodontist*, vol 78, no 4, pp. 761-765, 2008
- [15] E. Sicurezza, M. Greco, D. Giordano, F. Maiorana, R. Leonardi, "Accuracy of landmark identification on postero-anterior cephalograms," *Progress in Orthodontics*, in press
- [16] S. Cerutti, "Multivariate and multiscale analysis of biomedical signals: towards a comprehensive approach to medical diagnosis," *Proceedings of the 25th IEEE International Symposium on Computer-Based Medical Systems*, Rome, Italy, 2012.
- [17] R.C. Jacobson, "Facial analysis in two and three dimensions. In. (ed): *Radiographic cephalometry*," Chicago: Quintessence Publishing Co., Inc., 1995; 273-94.
- [18] P.J. Ford, K. Hibberd, "Creating effective and engaging information literacy programmes for the dental curriculum," *European Journal of Dental Education*, vol. 16, pp. e41-e46, 2012
- [19] J. Vuchkova, T. Maybury, C.S. Farah, "Digital interactive learning of oral radiographic anatomy," *European Journal of Dental Education*, vol. 16, pp. e79-e87, 2012
- [20] D. Murdoch-Eaton, "Feedback: the complexity of self-perception and the transition from 'transmit' to 'received and understood'," *Medical Education*, vol.46, pp. 538-540, 2012.
- [21] F. Maiorana, "A semantically enriched medical literature mining framework," *Proceedings of the 25th IEEE International Symposium on Computer-Based Medical Systems*, Rome, Italy, 2012.

Investigation of the Spatial Structure of Myomodulin E Molecule by Computer Modeling

N.A.Akhmedov, L.N.Agaeva*, R.M.Abbasli, L.I.Ismailova, N.M.Godjajev*

Institute for Physical Problems, Baku State University, Z. Khalilov Str.23, AZ-1148;
*Qafqaz University, Baku-Sumgayit road, 16th km. AZ0101, Khirdalan, Baku, Azerbaijan.
E-mail:Namig.49@bk.ru

Abstract-The spatial structure and conformational properties of the myomodulin E molecule have been investigated using methods of theoretical conformational analysis. Low-energy conformations of the heptapeptide molecule have been found and values of dihedral angles of main and side chains and energy of intra- and inter-residue interactions were estimated.

Key words: myomodulin peptide, conformation, structure, molecule

The myomodulin family of neuropeptides is an important group of neural cotransmitters in molluscs and is known to be present in the neural network that controls feeding behavior in the snail *Lymnaea* [1, 2]. Here they show that a single gene encodes five structurally similar forms of myomodulin: GLQMLRLamide, QIPMLRLamide, SMSMLRL

amide, SLSMLRLamide, and PMSMLRLamide, the latter being present in nine copies. Analysis of the organization of the gene indicates that it is transcribed as a single spliced transcript from an upstream promoter region that contains multiple cAMP-responsive elements, as well as putative elements with homology to tissue-specific promoter-binding sites. The presence in nervous tissue of two of the peptides, GLQMLRLamide and PMSMLRLamide, is confirmed by mass spectrometry. In situ hybridization analysis indicates that the gene is expressed in specific cells in all ganglia of the CNS of *Lymnaea*, which will allow physiological analysis of the function of myomodulins at the level of single identified neurons [1, 2].

TABLE I
THE RELATIVE ENERGY (U_{rel}) AND ENERGY CONTRIBUTIONS OF NONVALENT (U_{nv}), ELECTROSTATIC (U_{el}), TORSION (U_{tors}) INTERACTIONS OF OPTIMAL CONFORMATIONS OF GLY1-LEU2-GLN3-MET4-LEU5-ARG6-LEU7-NH₂ MOLECULE

№	Conformation	U_{rel}	Energy range, kcal/mol		
			U_{nv}	U_{el}	U_{tors}
1	BB ₁₁ B ₂₁₁ B ₂₁₂₂ R ₂₁ B ₃₃₂₂ R ₂₁	3.4	9.5	5.1	-21.9
2	RB ₂₁ R ₂₁₁ B ₂₁₂₂ R ₂₁ B ₃₃₂₂ R ₂₁	0	9.7	4.6	-25.3
3	RR ₂₁ R ₂₁₁ R ₂₂₂₂ B ₂₁ R ₃₁₂₂ R ₃₂	1.1	9.1	5.5	-24.2
4	BB ₁₂ R ₂₁₁ B ₂₁₂₂ R ₂₁ B ₃₃₂₂ R ₂₁	2.9	9.8	4.6	-22.4
5	RR ₁₂ B ₁₂₁ B ₂₁₂₂ R ₂₁ B ₃₂₂₂ R ₂₁	5.1	9.6	6.2	-20.2
6	RR ₂₁ R ₃₁₁ B ₂₁₂₂ B ₃₁ R ₃₁₂₂ R ₂₁	6.5	11.3	6.1	-19.8
7	RR ₂₁ R ₃₁₁ B ₂₃₂₂ B ₂₁ B ₃₃₂₂ B ₃₂	3.4	10.2	5.8	-21.9
8	RR ₂₁ R ₃₁₁ R ₂₂₂₂ B ₂₁ B ₃₃₂₂ B ₃₂	1.5	8.9	5.6	-23.8
9	BR ₂₁ R ₃₁₁ B ₂₂₂₂ R ₂₁ R ₃₃₂₂ R ₃₂	6.7	10.7	9.7	-18.6

Calculation of myomodulin E molecule has been carried out using methods of theoretical conformational analysis with regard to nonvalent, electrostatic and torsional interactions and energy of the hydrogen bonds. In the presentation of results of molecule calculations, one can use the classification suggested in the work [3-6]. According to it all structural versions break down into shapes including certain forms of the main chain and each form is represented by a set of conformations. The conformations are determined by the number of rotational degrees of freedom of the side chains of the residues being included in the molecule.

Notations and indications of dihedral angles correspond to generally accepted nomenclature [7].

Nonvalent interactions are estimated by Lennard-Jones potential with the parameters suggested in the work [8]. Electrostatic interactions are calculated in the monopole approximation by Coulombs law using partial charges on the atoms suggested in this work [8]. Conformational possibilities of cardio active peptides have been studied under the aqueous surroundings in connection with which the value of permittivity is taken to be 10 [9]. Hydrogen bonds are evaluated by Morze potential [9]. Torsional potentials and values of rotation barriers of amino acid main and side chain dihedral angles are taken from work [8].

TABLE II
ENERGY OF THE INTRA- AND INTER – RESIDUE INTERACTIONS (KCAL/MOL) IN CONFORMATIONS №2 ($U_{rel}=0$ KCAL/MOL, UPPER LINE), №3 ($U_{rel}=1.1$ KCAL/MOL, MIDDLE) AND №8 ($U_{rel}=1.5$ KCAL/MOL, LOWER LINE) OF THE GLY1-LEU2-GLN3-MET4-LEU5-ARG6-LEU7-NH₂ MOLECULE

Gly1	Leu2	Gln3	Met4	Leu5	Arg6	Leu7	
3.2	-1.3	-0.9	-3.7	-2.3	1.4	0	Gly1
3.6	-1.3	-1.3	-2.1	-2.3	1.3	0	
3.6	-1.3	-1.7	-2.3	-2.0	-1.5	0	
	-1.0	-1.4	-1.0	-0.2	-0.1	0	Leu2
	-1.0	-0.7	-1.1	-1.8	-3.2	-0.7	
	-1.1	-2.8	-1.2	-3.1	-0.3	-0.9	
		-0.7	-2.0	-0.2	0.1	-0.3	Gln3
		-0.8	-1.5	-0.9	-2.8	0	
		0.4	-0.9	-1.4	-1.7	0	
			0.6	1.6	-2.5	-4.3	Met4
			0.7	-1.2	-2.6	0	
			0.6	-1.0	-3.4	0	
				-1.0	-5.9	-1.4	Leu5
				-1.2	-2.7	-4.3	
				-0.6	-3.1	-3.3	
					1.4	-1.6	Arg6
					0.7	-0.1	
					0.8	-1.8	
						-3.3	Leu7
						-2.8	
						-2.7	

The conformational state of each amino residue is conveniently described by the backbone ϕ , ψ and side chain χ_1 , χ_2 ... dihedral angles. The terms “conformational state” or “conformation” used in the following analysis will always imply exact quantitative characteristics of residue or fragment geometry. For a stable conformation, the ϕ and ψ dihedral angles are located in low-energy region R, B, L and P of the conformational map. We introduce the notion “form of a residue” to denote the region of its backbone dihedral angle (R, B, L and P). The conformation of the backbone forms of residue in a given amino acid sequence will specify the backbone form of a fragment. The suggested notations are semi-quantitative characteristics of backbone geometries, which have no direct connection to the

actual order of amino acids in a sequence, but merely describe interactions of backbone elements, and also reflect potential side chain interactions with the backbone and with each other. All backbone forms of a dipeptide can be classified into two types, referred to as shapes: folded (f) and extended (e). The f-shape is represented by R-R, R-B and B-L forms, and the e-shape by B-B, B-R, R-L, L-B and L-R forms. Forms belonging to a particular shape have an analogous peptide chain contour and a similar mutual arrangement of backbones and side chains, and thus should exhibit similar medium-range interaction potentialities. Differences may arise from variations in short-range interactions and conformational freedom. A shape is an entirely qualitative category not related to any particular amino acid sequence.

TABLE III
GEOMETRIC PARAMETERS (DEGREE) OF THE OPTIMAL CONFORMATIONS OF GLY1-LEU2-GLN3-MET4-LEU5-ARG6-LEU7-NH₂ MOLECULE
(the values of dihedral angles are given in the sequence ϕ , ψ , ω , χ^1 , χ^2 ,...)

Residues	Conformation								
	RB ₂₁ R ₂₁₁ B ₂₁₂₂ R ₂₁ B ₃₃₂₂ R ₂₁			RR ₂₁ R ₂₁₁ R ₂₂₂₂ B ₂₁ R ₃₁₂₂ R ₂₂			RR ₂₁ R ₃₁₁ R ₂₂₂₂ B ₂₁ B ₃₃₂₂ B ₃₂		
Gly1	-72	-72	-178	-84	-81	177	-83	-76	177
Leu2	-70	-49	174	-69	-39	178	-70	-35	176
	176	62	179	176	62	179	178	61	179
	176			176			175		
Gln3	-60	-51	179	-59	-46	-179	-65	-42	177
	180	63	87	180	62	86	-78	63	81
Met4	-159	125	-178	-70	-57	179	-68	-36	180
	-179	62	180	179	177	180	-176	176	180
	-179			-177			180		
Leu5	-71	-60	-178	-115	118	-174	-71	128	-176
	174	66	179	180	65	179	178	65	-178
	177			175			176		
Arg6	-96	72	-174	-71	-31	179	-104	113	-179
	-50	-59	179	-81	63	174	-47	-61	173
	175			-179			178		
Leu7	-89	-59	180	-55	-46	-179	-111	120	180
	-179	62	179	175	63	180	-55	174	-172
	175			176			180		
U_{rel} (kcal/mol)	0			1.1			1.5		

To designate conformational states of the residues we have used X (*i, j*) - typed identifiers, where X defines low-energy regions of the conformational map φ - ψ : R ($\varphi, \psi = -180-0^\circ$), B ($\varphi = -180-0^\circ, \psi = 0-180^\circ$), L ($\varphi, \psi = 0-180^\circ$), and P ($\varphi = 0-180^\circ, \psi = -180-0^\circ$), *i, j*...=11...,12...,13...,21..., etc. conform to the positions of the side chain (χ_1, χ_2 ...), subscript 1 corresponds to the angle $\chi = 0-120^\circ$; 2 to $\chi = 120-(-120)^\circ$, 3 to $\chi = (-120)-0^\circ$.

Spatial structure of heptapeptide molecule has been investigated in fragments. The conformational possibilities of the N-terminal Gly1-Leu2-Gln3-Met4 tetrapeptide have been studied at the first step on the base of low energy conformations of appropriate amino acid residues. Calculation of Met4-Leu5-Arg6-Leu7-NH₂ fragment spatial structure is carried out on the base of stable conformations of amino acid residues. At the final stage calculation of tetrapeptides enable us to evaluate the conformational properties of the whole molecule.

Half of more than 200 structural versions of heptapeptide molecule examined in this paper appear to be sterically prohibited; relative energy of other conformations varied within 0-40 kcal/mol range. The best optimal conformations of this molecule whose energy does not exceed 9kcal/mol have been presented in the Table I. They have 9 different forms of main chain which belong to 9 shapes of this molecule. There are the energy contributions of nonvalent (energy of hydrogen bonds contribution included in value U_{nv}), electrostatic, and torsion interactions energy in this table. For the most low-energy conformation of the heptapeptide molecule in Table II, energy of intra- and inter- residue interactions for three conformations are presented. In Table III, values of dihedral angles of main and side chains of these conformations are given.

The global conformation of heptapeptide molecule Gly1-Leu2-Gln3-Met4-Leu5-Arg6-Leu7-NH₂ is RB₂₁R₂₁₁B₂₁₂₂R₂₁B₃₃₂₂R₂₁. This conformation is efficient both in nonvalent and in electrostatic interactions forming hydrogen bonds between atoms of main chain which make contribution (-4.2 kcal/mol) in total energy. Theoretical conformational analysis of molecule Gly1-Leu2-Gln3-Met4-Leu5-Arg6-Leu7-NH₂ brings about such structural organizations of molecule which do not exclude realization by hormones for a variety of functions requiring strict specific interactions with different receptors.

This work has been fulfilled in the frame of collaboration treaty of the Qafqaz and Baku State Universities.

REFERENCES

- [1] Kellett E, Perry SJ, Santama N, Worster BM, Benjamin PR, Burke JF, "Myomodulin gene of Lymnala: structure, expression and analysis", *J Neurosci.*, vol. 16, N16, pp. 4949-4957, 1996.
- [2] Orekhova IV, Alexeeva V, Church PJ, Weiss KR, Brezina V. "Multiple presynaptic and postsynaptic sites of inhibitory modulation by myomodulin of ARC neuromuscular functions of Aplysia", *Neurophysiol.*, vol. 89, N3, pp. 1488-1502, 2003.
- [3] Popov E.M. "Quantity approach to conformations of proteins", *Int. J. Quant. Chem.*, vol.16, pp. 707-737, 1979.
- [4] Akhmedov N.A., Akhverdiyeva G.A., Godjaev N.M., Popov E.M. "Theoretical conformational analysis of brain peptides δ -Melanocyte-stimulating hormone", *Int. J. Peptide and Protein Res.*, vol. 27, pp. 95-111, 1986.
- [5] Ismailova L.I., Akhmedov N.A., Abbasli R.M., Spatial Structure of Izoleucine Pentapeptides Glu-Phe-Leu-Arg-Ile-NH₂ and Pro-Phe-Tyr-Arg-Ile-NH₂" *Biophysics*, vol.53, pp.14-21, 2008.
- [6] Ismailova L.I., Abbasli R.M., Akhmedov N.A. "Spatial structure of the cardio active oktapeptide", *Biophysics*, vol. 52: pp. 1141-1147, 2007.
- [7] IUPAC-IUB, "Joint Commission on Biochemical Nomenclature" *Blackwell Scientific Publications, Oxford*, vol. 39, 1988.
- [8] Momany F.A., McGuire R.F., Burgess A.W., Scheraga H.A. "Energy parameters in polypeptides. VII. Geometric parameters, partial atomic charges, nonbonded interactions, hydrogen bond interaction occurring amino acids" *J.Phys.Chem.*, vol.79, pp. 2361-2381, 1975.
- [9] Lipkind G.M., Archipova S.F., Popov E.M. "Theoretical conformational analysis of methylamides of N-asetil-L-amides", *J. Struct. Chem*, vol. 11, pp. 121-126, 1970.

Multi-objective Evolutionary Algorithm Based on Decomposition for Efficient Coverage Control in Mobile Sensor Networks

Bara'a Ali ATTEA
Computer Eng. Dept.
University of Baghdad
Baghdad / IRAQ
baraali@yahoo.com

Feyza Yıldırım OKAY
Computer Eng. Dept.
Gazi University
Ankara/TURKEY
feyzaokey@gazi.edu.tr

Suat ÖZDEMİR
Computer Eng. Dept.
Gazi University
Ankara/TURKEY
suatozdemir@gazi.edu.tr

M. Ali AKCAYOL
Computer Eng. Dept.
Gazi University
Ankara/TURKEY
akcayol@gazi.edu.tr

Abstract— Node deployment is a fundamental issue to be solved in Wireless Sensor Networks (WSNs). A proper node deployment scheme can reduce the complexity of problems in WSNs such as routing, data fusion, communication, etc. Furthermore, it can extend the lifetime of WSNs by minimizing energy consumption due to sensing and communication. This paper presents the development of a recently multi-objective optimization algorithm, the so-called multi-objective evolutionary algorithm based on decomposition (MOEA/D). MOEA/D aims the relocation of mobile nodes in a WSN with the goal of providing maximum sensing coverage area, and with the constraint of minimizing the energy required for the relocation as measured by total travel distance of the sensors from their initial locations to their final locations. Simulation results clearly show that the non-dominated solutions have tradeoff between the travelled distance and coverage area.

Keywords— Coverage control, energy conservation, evolutionary algorithm, mobile sensor network, multi-objective optimization

I. INTRODUCTION

A wireless sensor network (WSN) consists of small sensing devices that can be readily deployed in diverse environments to form a distributed wireless network for collecting information in a robust and autonomous manner. Possible applications of sensor networks are of interest to the most diverse fields. Environmental monitoring, warfare, child education, surveillance, micro-surgery, and agriculture are only a few examples.

A key challenge in WSNs is to determine a sensor field architecture that minimizes cost, provides high sensor coverage, resilience to sensor failure, and appropriate computation and communication tradeoff. Intelligent sensor deployment facilitates the unified design and operation of sensor systems, and decreases the need for excessive network communication.

A mobile sensor network (MWSNs) is composed of a distributed collection of nodes, each of which has sensing, computation, communication and locomotion capabilities. It is this latter capability that distinguishes a mobile sensor network from its more conventional static cousins. Locomotion facilitates a number of useful network capabilities, including the ability to self-deploy; that is, starting from some compact initial configuration, the nodes in the network can spread out such that the area 'covered' by the network is maximized [1].

For random deployment of sensors, past research has focused on how to move the sensors to positions that yield maximum coverage. Howard, *et al.*, [1] propose a potential-field-based approach to deployment. The fields are constructed such that each node is repelled by both obstacles and by other nodes, thereby forcing the network to spread itself throughout the environment. In [2], the authors propose a virtual force algorithm (VFA) as a deployment strategy for randomly-placed sensors. In VFA, a judicious combination of attractive and repulsive forces is used to determine virtual motion paths and the rate of movement. Once the effective sensor positions are identified, a one-time movement with energy consideration incorporated is carried out. They also propose a probabilistic target localization algorithm that is executed by the cluster head. The localization results are used by the cluster head to query only a few sensors (out of those that report the presence of a target) for more detailed information. Based on Voronoi diagrams, the authors in [3] propose two sets of distributed protocols for controlling the movement of sensors to achieve target coverage, one favoring communication and one favoring movement. In the first protocol, sensor nodes move iteratively, eventually reaching the final destination. In each iteration, sensor nodes detect coverage holes using a Voronoi diagram. If holes exist, they calculate the target locations to heal the holes and move. In the second one, virtual movement protocols, sensor nodes do not perform iterative physical movement. Instead, after calculating the target locations, sensor nodes move virtually and exchange these new virtual locations with the sensors which would be their neighbors if they had actually moved. The real movement only occurs when the communication cost to reach their logical neighbors is too high or when they determine their final destinations. In [4], the authors propose a distributed and scalable solution to the deployment problem of mobile sensor networks in unknown environments. Their approach is based on fluid dynamics through which we model the sensor network as a fluid body and each sensor node as a fluid element. Originating from local neighborhood interactions, the principles of fluid flow are adapted to the deployment of the sensor network. Using the concept of molecule spreading from physics, [5] presents Self-Deployment by Density Control (SDDC) as an efficient method for sensor deployment, assuming that global information is not available. SDDC uses density control by each node to concurrently deploy sensor nodes.

In this paper, we develop a new framework based on multi-objective evolutionary algorithm (MOEA) to determine

the sensor node deployment in MWSNs by minimizing energy consumption and maximizing the network coverage task. The paper is organized as follows. In the next section we give an overview of the concept of multi-objective evolutionary algorithm, concentrating on recent one, the so-called multi-objective evolutionary algorithm based on decomposition (MOEA/D). In Section 3, we formulate MOEA/D for efficient coverage control in MWSNs. Section 4 contains experimental results, and finally, Section 5 gives a brief discussion of conclusion and some future works.

II. MULTI-OBJECTIVE EVOLUTIONARY ALGORITHM BASED ON DECOMPOSITION

The task of obtaining the optimal solution in an optimization problem which involves only one objective function is called single objective optimization problem, however, many real-world optimization problems involve multiple objectives. A multi-objective optimization problem (MOP) can be formulated as follows:

$$\begin{aligned} \text{maximize } F(x) &= (f_1(x), \dots, f_n(x))^T \\ \text{subject to } x &\in \Omega \end{aligned} \quad (1)$$

where x is the decision variable vector, $F: \Omega \rightarrow \mathbb{R}^n$ consists of n real-valued objective functions, the objective space is \mathbb{R}^n and Ω is the search space. In general, f_1, \dots, f_n are in conflict with each other, and then finding the optimum can be interpreted as finding a good trade-off between all f_1, \dots, f_n of F . Thus, the multi-objective approach to solving MOPs generates “partial orders” of solutions leading to possible multitudes of trade-off solutions in objective space.

While for single-objective optimization, a “total order” exists, MOP can then be solved using an aggregated weight sum of all the objectives. A popular scalarizing approach for the aggregated weight sum is the linear fitness combination technique using (considering maximization problem):

$$\text{Fitness} = \max \sum_{i=1}^n w_i f_i(x) \quad (2)$$

where $w_i \geq 0$ and $i = 1, \dots, n$ are the weighting coefficients representing the relative importance of the i^{th} objective functions of the MOP. It is usually assumed for normalization that:

$$\sum_{i=1}^n w_i = 1 \quad (3)$$

On the other hand, a multi-objective optimization (MOO) algorithm can be implied to find the non-dominated points (Pareto front) [6] [7]. Suppose x and y are two decision variable vectors, x is said to *dominate* y , denoted by $x \succ y$, if and only if $f_i(x) \geq f_i(y)$ for every $i \in \{1, 2, \dots, n\}$ and $f_j(x) > f_j(y)$ for at least one index $j \in \{1, 2, \dots, n\}$. Then, $x^* \in \Omega$ is said to be non-dominated if there is no other $x \in \Omega$ so that $F(x)$ dominates $F(x^*)$. Figure 1 shows a particular case of the dominance relation in the presence of two objective functions [7].

There are several variants of multi-objective evolutionary algorithms (MOEAs), but with the common aim of how to retain the non-dominated solutions generated during the search. Recent MOEAs can be classified into a general scheme [7]. First, the initial population is filled by randomly generated solutions. Next, given the objective functions to be maximized, the population is evaluated and ranked on the basis of non-domination and distribution. Based on this rank, some of the best candidates are chosen to seed the next

generation by applying variation (i.e. crossover and mutation) operators. Next, the non-dominated solutions from both, the parents and offspring population are merged to become the parents for the next population. This process is repeated until a stop condition is reached.

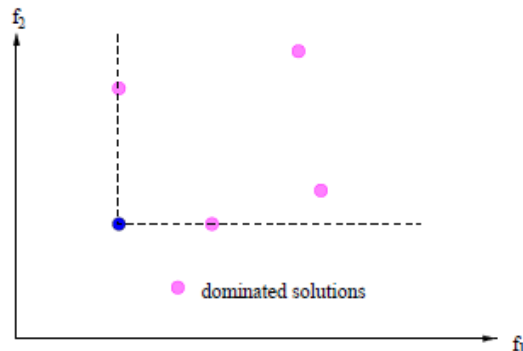


Figure 1. Dominance relation in a bi-objective space.

The main drawback of the generic evolutionary multi-objective techniques is that it treats a MOP as a “black box”, i.e. without using problem-specific knowledge, which may have undesirable effects, such as forcing the evolutionary process into unnecessary searches and destructive mating, negatively affecting their overall performance. Therefore, the incorporation of problem-specific knowledge in MOEAs to direct the search into promising areas of the search space can be proven beneficial. However, designing problem-specific operators for a multi-objective optimization problem (MOP), as a whole, is difficult. The Multi-Objective Evolutionary Algorithm based on Decomposition (MOEA/D) [8] alleviates this difficulty by decomposing the MOP into many scalar sub-problems that are optimized in parallel, by using neighborhood information and scalar techniques. At each generation, the population is composed of the best solution found so far (i.e. since the start of the run of the algorithm) for each sub-problem. The neighborhood relations among these sub-problems are defined based on the distances between their aggregation coefficient vectors. The optimal solutions to two neighboring sub-problems should be very similar. Each sub-problem (i.e., scalar aggregation function) is optimized in MOEA/D by using information only from its neighboring sub-problems. The general framework of MOEA/D can be presented in what follow [8]. Let $\lambda^1, \dots, \lambda^N$ be a set of even spread weight vectors and $z^* = (z_1^*, \dots, z_n^*)$ be the reference point. The problem of approximation of the PF of the MOP in Eqn. 1 can be decomposed into scalar optimization sub-problems using the Tchebycheff approach and the single objective optimization j^{th} sub-problem in this approach is [9]:

$$\text{minimize } g^{\text{te}}(x|\lambda^j, z^*) = \max\{\lambda_i^j |f_i(x) - z_i^*|\} \quad (4)$$

$$1 \leq i \leq n$$

where $\lambda^j = (\lambda_1^j, \dots, \lambda_n^j)^T$ is a weight vector, i.e., $\forall i = 1, \dots, n: \lambda_i \geq 0$ and $\sum_{i=1}^n \lambda_i = 1$. MOEA/D optimizes all these N objective functions simultaneously in a single run. At each generation t , MOEA/D with the Tchebycheff approach maintains: a population of N points $x^1, \dots, x^N \in \Omega$, where x^i is the current solution to the i^{th} subproblem, FV^1, \dots, FV^N where $FV^i = F(x^i) \forall i = 1, \dots, N$; and $z = (z_1, \dots, z_n)^T$ where z_i be the best value found so far for objective f_i . Also, MOEA/D

maintains an external population (EP), which is used to as an archive strategy to store the non-dominated solutions found during the search. Another commonly used decomposition approach is the Weighted Sum Approach [10]. Let $\lambda = (\lambda_1, \dots, \lambda_n)^T$ be a weight vector, i.e., $\forall i = 1, \dots, n: \lambda_i \geq 0$ and $\sum_{i=1}^n \lambda_i = 1$. Then, the optimal solutions to the following single optimization problems:

$$\begin{aligned} \text{maximize } g^{ws}(x|\lambda) &= \sum_{i=1}^n \lambda_i f_i(x) \\ \text{subject to } x &\in \prod_{i=1}^n [a_i, b_i] \end{aligned} \quad (5)$$

Are Pareto optimal to Eqn. 1 if the PF of (1) is convex, where $g^{ws}(x|\lambda)$ is used to emphasize that λ is a weight vector in this objective function. However, when the Pareto set (PS) is not convex, the weighted sum approach could be not able to find some Pareto optimal solutions.

The general framework of MOEA/D can proceed as in Algorithm 1 [8].

Algorithm 2. The MOEA/D general framework

Input:

- MOP
- the number of the sub-problems considered in MOEA/D, N
- a uniform spread of N weight vectors: $\lambda^1, \dots, \lambda^N$
- the number of the weight vectors in the neighborhood of each weight vector, T
- the maximum number of generations, gen_{max}

Output:

- EP

Step 0 - Setup:

- Set $EP = \emptyset$
- $gen = 0$

Step 1 – Initialization

- Uniformly randomly generate an initial internal population, $IP_0 = \{x^1, \dots, x^N\}$ and set $FV^i = F(x^i)$.
- Initialize $z = (z_1, \dots, z_n)^T$ by a problem-specific method.
- Compute the Euclidean distances between any two weight vectors and then work out the T closest weight vectors to each weight vector. $\forall i = 1, \dots, N$, set $B(i) = \{i_1, \dots, i_T\}$, where $\lambda^{i_1}, \dots, \lambda^{i_T}$ are the T closest weight vectors to λ^i .

Step 2 – Update: For $i = 1, \dots, N$

- Genetic operators: Randomly select two indexes k, l from $B(i)$, and then generate a new solution y from x^k and x^l by using genetic operators.
- Update of z , $\forall j = 1, \dots, n$, if $z_j < f_j(y)$, then set $z_j = f_j(y)$.
- Update of Neighboring Solutions: For each index $j \in B(i)$, if $g^{te}(y|\lambda^j, z) \leq g^{te}(x^j|\lambda^j, z^*)$, then set $x^j = y$ and $FV^j = F(y^j)$.
- Update of EP: Remove from EP all the vectors dominated by $F(y)$. Add $F(y)$ to EP if no vector in EP dominate $F(y)$.

Step 3 – Stopping criteria

- If $gen = gen_{max}$, then stop and output EP, otherwise $gen = gen + 1$, go to Step 2.

III. MOEA/D FOR COVERAGE EFFICIENCY IN MWSNs

This section presents how MOEA/D addresses the energy conservation and area coverage design issues for MWSNs. We model these two design issues as a multi-objective optimization problem using the following definition. Let \mathcal{A} be a square monitoring field with known dimensions, \mathcal{BS} be the base-station or sink node with its coordinates $(x_{BS}, y_{BS}) \in \mathcal{A}$, $\mathcal{S}_{\theta_S} = \{s_1, \dots, s_k\}$ be a set of k sensor nodes with set Θ_S of parameters concerning the initial spatial locations, coverage

radii, and initial energies of all the sensor nodes defined in Eqn. (5) (In our simulation, we assume that each sensor's coverage is a circle. Moreover, we assume that all sensors have same coverage with radius r_s . This assumption simplifies our simulations. However, sensors with different coverage can be also used by our algorithm.) Also, let $\mathcal{D}_{\theta_D} = \{d_1, \dots, d_l\}$ be a set of l demand points with set $\Theta_D = \{(x_{d_1}, y_{d_1}), \dots, (x_{d_l}, y_{d_l})\}$ of their spatial locations.

$$\Theta_S = \{(x_{s_1}, y_{s_1}, r_{s_1}, E_{s_1}), \dots, (x_{s_k}, y_{s_k}, r_{s_k}, E_{s_k})\} \quad (6)$$

The coverage efficiency problem in a MWSN consists of ensuring that each demand point d of a fraction $\mathcal{C} \subset \mathcal{D}$ is covered by at least one sensor node $s \in \mathcal{S}$ but with the interest of minimum travelled distance problem (i.e., the final positions of the sensors and the travelled distance are optimized in order to get the best coverage in the most energy efficient way.) In this paper, we address this problem as a multi-objective problem with the goal of minimizing the total travelled distance, D and increasing area coverage, \mathcal{C} . We address this tradeoff using MOEA/D for obtaining multiple non-dominated solutions indicating the tradeoff between the two objectives. The following formulations illustrate the characteristic components of MOEA/D for efficient area coverage problem in MWSNs. Let MOEA/D to be defined as an 8-tuple:

$$MOEA/D = (I, \Phi, \Omega, \Psi, \iota, N, EP, \varphi) \quad (7)$$

where I is the individual space being encoded as a complete solution. Each individual is represented as a set of k Cartesian sensor locations, as shown in Eqn. 7. Thus, I is a chromosome and each 2D Cartesian coordinate, which represents a sensor location, is a gene.

$$I = \{(x_{s_1}', y_{s_1}'), (x_{s_2}', y_{s_2}'), \dots, (x_{s_k}', y_{s_k}')\} \quad (8)$$

Then, an internal population $IP = \{I_1, \dots, I_N\} \in I^N$, of N individual solutions can be formally specified as:

$$\begin{aligned} \forall i \in \{1, \dots, N\} \text{ and } \forall j \in \{1, \dots, k\} \\ I_{i,j} = (x_{s_j}, y_{s_j}) \end{aligned} \quad (9)$$

The first step of the algorithm is to generate an initial population. This is usually done randomly. The second step involves the calculation of the fitness vector for each individual in the population, IP . Let $\Phi: I \rightarrow \mathbb{R}^2$ denotes the fitness (i.e., objective) function vector assigning total travelled distance, D , and non-coverage penalty, NC , to individuals.

$$\begin{aligned} \forall i: 1 \leq i \leq k \\ \Phi_i = \{D_i, NC_i\} \end{aligned} \quad (10)$$

D is then calculated as in Eqn. 10

$$D(I) = \frac{\sum_{i=1}^k |C_i - C_i'|}{k} = \frac{\sum_{i=1}^k \sqrt{(x_i - x_i')^2 + (y_i - y_i')^2}}{k} \quad (11)$$

where C_i and C_i' be the corresponding initial and final Cartesian coordinate of sensor node i .

NC is defined as minimizing the number of uncovered target points:

$$NC(I) = \sum_{i=1}^m Uncovered(d_i) \quad (12)$$

Where:

$$\begin{aligned} Uncovered(t_i) = \\ \begin{cases} 0 & \text{if } \exists s \in \mathcal{S} \text{ and } d(s, t_i) \leq r_s \\ 1 & \text{otherwise} \end{cases} \end{aligned} \quad (13)$$

Ω is the set of genetic operators: selection, crossover, and mutation, each of which is controlled by specific parameters summarized by Θ :

$$\Omega = \{s_{\Theta_s}, c_{\Theta_c}, m_{\Theta_m} \mid s_{\Theta_s}, c_{\Theta_c}, m_{\Theta_m} : I^N \rightarrow I^N\} \quad (14)$$

Crossover and mutation are the perturbation operators, which can alter the individual solutions found in the population. In our problem the variables that are needed to be optimized are the Cartesian coordinates of the sensors. Therefore, the input variables to the genetic algorithm are as the coordinates of the sensor nodes. A proportion p_c of pairs of parents in the population are chosen for crossover. Once a partner is chosen, a random number of genes (sensor locations $r1, r2$), chosen at random, in the range $\{1, \dots, k-1\}$, are exchanged, resulting in two children as presented in Eqn. 14. Figure 2 illustrates an example of crossover for two individual solutions, each of ten genes. In the example, $r1 = 5$ and $r2 = 7$.

$$c_{p_c} : I^2 \rightarrow I^2$$

$$I'_1 = (I_{1,1}, \dots, I_{1,r1}, I_{2,r1+1}, \dots, I_{2,r2}, I_{1,r2+1}, \dots, I_{1,k})$$

$$I'_2 = (I_{2,1}, \dots, I_{2,r1}, I_{1,r1+1}, \dots, I_{1,r2}, I_{2,r2+1}, \dots, I_{2,k}) \quad (15)$$

Parent1									
S_1	S_2	S_3	S_4	S_5	S_6	S_7	S_8	S_9	S_{10}
x	y	x	y	x	y	x	y	x	y
S_1	S_2	S_3	S_4	S_5	S_6	S_7	S_8	S_9	S_{10}
x	y	x	y	x	y	x	y	x	y
Parent2									
S_1	S_2	S_3	S_4	S_5	S_6	S_7	S_8	S_9	S_{10}
x	y	x	y	x	y	x	y	x	y
S_1	S_2	S_3	S_4	S_5	S_6	S_7	S_8	S_9	S_{10}
x	y	x	y	x	y	x	y	x	y
Child1									
S_1	S_2	S_3	S_4	S_5	S_6	S_7	S_8	S_9	S_{10}
x	y	x	y	x	y	x	y	x	y
S_1	S_2	S_3	S_4	S_5	S_6	S_7	S_8	S_9	S_{10}
x	y	x	y	x	y	x	y	x	y
Child2									
S_1	S_2	S_3	S_4	S_5	S_6	S_7	S_8	S_9	S_{10}
x	y	x	y	x	y	x	y	x	y

Figure 2. An example of 2-point crossover.

Each allele in the new individuals is then mutated with the probability p_m . A biased coin is flipped for each gene in all chromosomes, where the probability of heads is equal to p_m and tails is $1 - p_m$. When the coin-flip results in heads, a new random sensor location is chosen, as in Eqn. 15, for the gene and the altered chromosome becomes a new population member. A sensor's location remains unchanged when a coin-flip results in tails. Figure 3 illustrates an example of mutation operation, where mutation occurs at gene S_4 .

$$I'_{ij} = (x_{s_j}', y_{s_j}') \mid x_{s_j}' \text{ and } y_{s_j}' \in \mathcal{A} \quad (16)$$

Child									
S_1	S_2	S_3	S_4	S_5	S_6	S_7	S_8	S_9	S_{10}
x	y	x	y	x	y	x	y	x	y
S_1	S_2	S_3	S_4	S_5	S_6	S_7	S_8	S_9	S_{10}
x	y	x	y	x	y	x	y	x	y
Mutated Child									
S_1	S_2	S_3	S_4	S_5	S_6	S_7	S_8	S_9	S_{10}
x	y	x	y	x	y	x	y	x	y

Figure 3. An example of mutation.

The generated child can be used to update the neighborhood of the current sub-problem. Also, this child may

update the status of EP . The generation EP update function $\Psi: EP \rightarrow EP'$ described the process of updating the current EP by removing and/or adding dominated and/or non-dominated solutions while applying Ω to the current I^N . Finally, $t: I^N \rightarrow \{true, false\}$ is a termination criterion for $MOEA/D$. A decision maker function $\varphi: EP \rightarrow I^*$ can select one individual solution I^* out of EP (e.g., to select the solution which expends the minimum travelled distance for ensuring a coverage level greater than or equal to the minimum acceptable one) and decodes it into an explicit MWSN design.

IV. SIMULATION RESULTS

The simulations are performed on 10 different wireless sensor networks. Each WSN is composed of 20 sensor nodes and 50 target points deployed randomly in a playground of $100m \times 100m$ sensor field (much of the WSN literature assumes that the sensors will be randomly deployed). This means that the horizontal and vertical coordinates of each sensor and target are randomly selected between zero and the maximum value of the dimension (however, the initial Cartesian coordinates of the sensors were selected to be in one corner of the network area \mathcal{A} , so as to see the ability of $MOEA/D$ to scatter their locations into the whole area of \mathcal{A}). We assume a center-located BS, i.e., the maximum distance of any node from the BS is about $70m$. The evolutionary components of $MOEA/D$ are fixed to binary tournament selection, two-point crossover with $p_c = 0.6$, and mutation with $p_m = 0.03$, respectively. The population size, N , is taken as 100 and allowed to evolve for 50 generations. Neighborhood size, T , is selected to be 4. Due to space limitations, we are able to present the results for only two networks as follows. Complete results will take place in the full version of this paper. Interested readers are more than welcome to contact authors for further results.

Figure 4 depicts two different randomly created WSNs, along with their initial sensors and targets locations and the total number of initial detected targets. Undetected targets are drawn in yellow plus sign while those detected in blue plus ones.

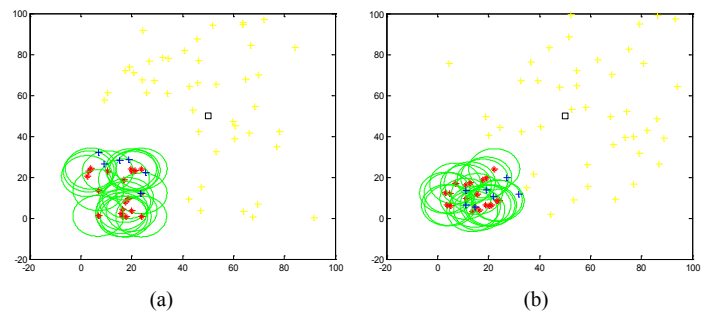


Figure 4. Initial WSNs (a) WSN#1 the number of initial detected targets is 6, (b) WSN#2 the number of initial detected targets is 7.

Figure 5 shows the evolution of a WSN after every 10 generations. The depicted solution presents one of the non-dominated solutions extracted from the EP set. EP set, exhibit some form of evolution after each 10 generations.

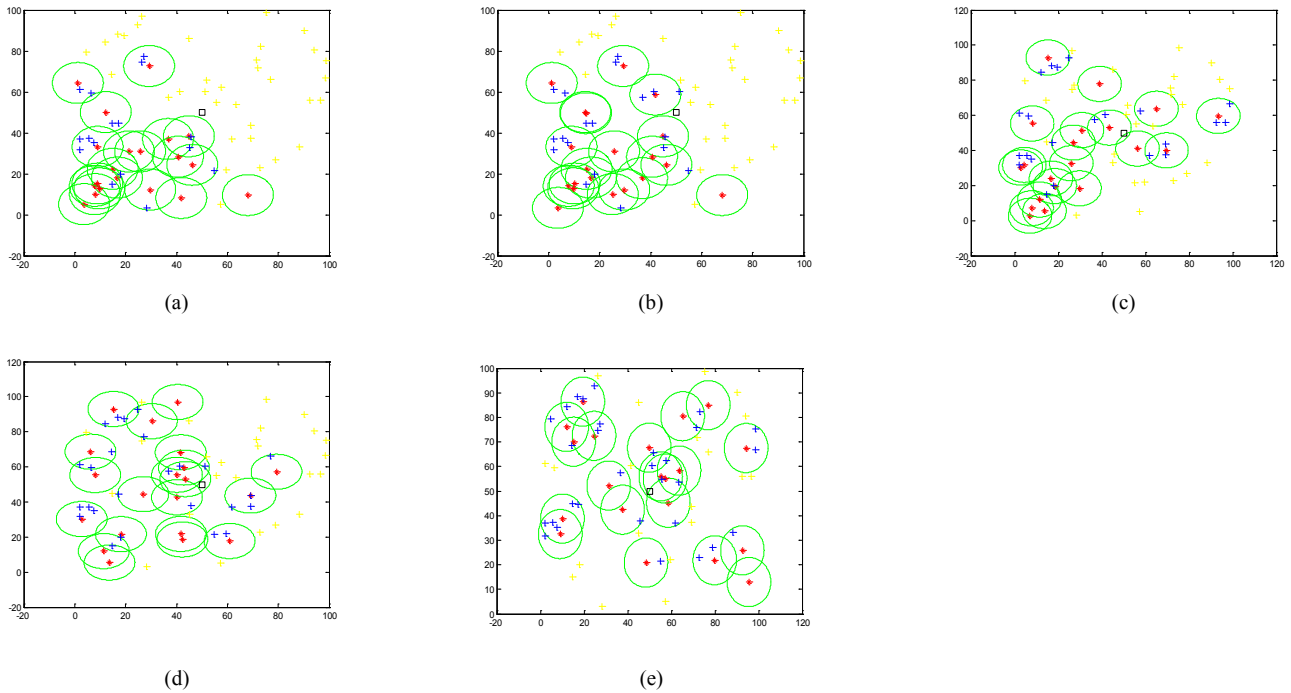


Figure 5. Non-dominated solutions for every 10 generations (a) After 10th generation with 16 detected targets, (b) After 20nd generation with 19 detected targets, (c) After 30rd generation with 22 detected targets, (d) After 40th generation with 25 detected targets, (e) After 50th generation with 30 detected targets.

In Figure 6 we present seven different non-dominated solutions extracted from the evolved EP after 50 generations. Each solution is depicted as WSN with its corresponding Fitness1 and Fitness2 functions.

Moreover, Figure 7 presents the EP non-dominated solutions as a graph which clearly represents the trade of between the Fitness functions.

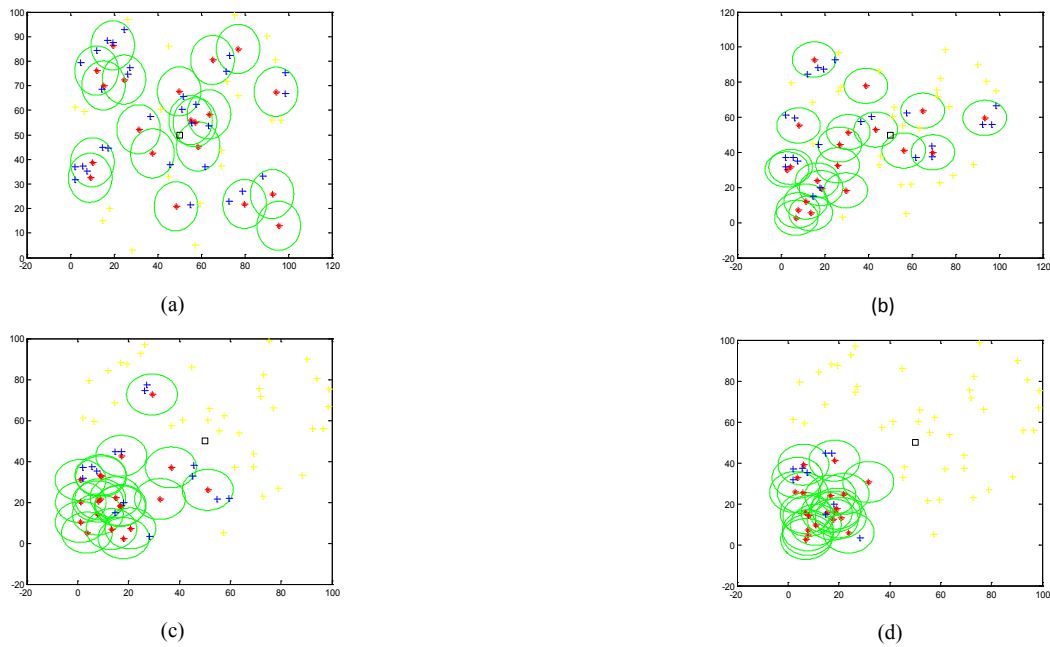


Figure 6. EP solution set after 50. generations (a) EP solution #1. (b) EP solution #2. (c) EP solution #3. (d) EP solution #4

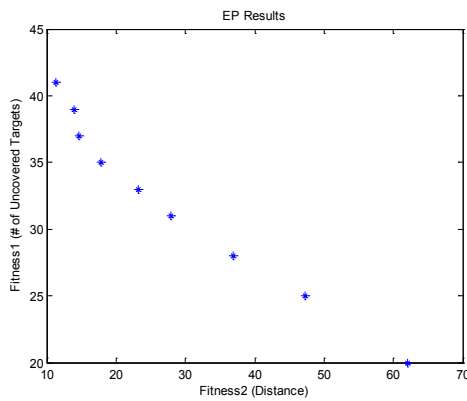


Figure 7. EP Non-dominated solutions as a graph

TABLE I: QUANTITATIVE RESULTS OF EP SOLUTIONS AFTER EVERY 10 GENERATIONS WITH A TOTAL OF 50 GENERATIONS

	Fitness 1 (Number of Undetected Targets)	Fitness 2 (Distance)
EP Set # 1	24	43,7081
	27	32,2871
	29	28,4248
	31	26,6933
	36	25,6783
	37	24,2335
	39	21,9261
EP Set # 2	41	21,6770
	24	43,7081
	27	32,2871
	29	28,4248
	31	26,6933
	36	25,6783
	37	24,2335
EP Set # 3	38	19,3071
	39	13,2695
	24	43,7081
	27	32,2871
	29	28,4248
	31	26,6933
	36	25,6783
EP Set # 4	37	24,2335
	38	19,3071
	39	12,9929
	24	43,7081
	27	32,2871
	29	28,4248
	31	26,6933
EP Set # 5	36	25,6783
	37	24,2335
	38	19,3071
	39	12,9929
	40	12,2537
	42	11,1732
	44	10,8589
EP Set # 5	24	43,7081
	27	32,2871
	29	28,4248
	31	26,6933
	36	25,6783
	37	24,2335
	38	19,3071
EP Set # 5	39	12,9929
	40	12,2537
	41	10,0982
	42	9,3777

Finally, Table I quantitatively presents the evolved EP after every 10 generations. Each EP has different number of non-dominated solutions. It can be seen from the Table I that solutions get better in every generation.

V. CONCLUSION

This paper introduces a MOEA/D for relocating mobile nodes of a MWSN in order to provide the trade-off between coverage and nodes' travelled distance. Both the coverage and total distance are considered in the optimization. MOEA/D doesn't offer only one best solution, so we have a solution set. After determining the objective functions (i.e. Fitness functions), MOEA/D will evolve a set of initial solutions to a number of non-dominated solutions, each has a bias toward one of the two fitness functions. Also, a decision maker for the non-dominated solutions of the EP set can be used up to our needs.

ACKNOWLEDGMENT

This work is supported in part by the Gazi University Scientific Research Project Funds No. 06/2011-09 and 06/2012-49.

REFERENCES

- [1] A. Howard, M. J. Matari'c, and G. S. Sukhatme, "Mobile Sensor Network Deployment using Potential Fields: A Distributed, Scalable Solution to the Area Coverage Problem," In *Proceedings of the 6th International Symposium on Distributed Autonomous Robotics Systems (DARS02)*, pp. 299-308, 2002.
- [2] Yi Zou, K. Chakrabarty, "Sensor Deployment and Target Localization Based on Virtual Forces", *Proc. IEEE INFOCOM*, vol 2 pp:1293-1302, 2003.
- [3] Guiling Wang, Guohong Cao, T. F. Porta, "Movement-Assisted Sensor Deployment", *IEEE INFOCOM*, vol 4, pp: 2469-2479, 2004.
- [4] Muhammad R. Pac, Aydan M. Erkmn, Ismet Erkmn, "Scalable Self-Deployment of Mobile Sensor Networks: A Fluid Dynamics Approach," *Proceeding of 2006 IEEE/RSJ International Conference on intelligent Robots and Systems*, pp: 1446-1451, 2006.
- [5] Ruay-Shiung Chang, Shuo-Hung Wang, "Self-Deployment by Density Control in Sensor Networks," *IEEE transactions on vehicular technology*, vol. 57, no.3, pp 1745-1755, 2008.
- [6] K. Deb, *Multi-Objective Optimization Using Evolutionary Algorithms*. Wiley and Sons, 2002.
- [7] C.A.C. Coello, G.B. Lamont, and D.A. Van Veldhuizen, *Evolutionary Algorithms for Solving Multi-Objective Problems*, 2nd Edition, Springer, 2007.
- [8] Q. Zhang and H. Li, MOEA/D: a multi-objective evolutionary algorithm based on decomposition, *IEEE Transactions on Evolutionary Computation* 11 (6) (2007) 712-731.
- [9] H. Li and Q. Zhang, " Multiobjective Optimization Problems With Complicated Pareto Sets, MOEA/D and NSGA-II", *IEEE Transactions On Evolutionary Computation*, Vol. 13, NO. 2, pp. 284-302, Apr. 2009.
- [10] Q. Zhang, W. Liu, E. Tsang, and B. Virginas, "Expensive Multiobjective Optimization by MOEA/D with Gaussian Process Model", *IEEE Transactions On Evolutionary Computation*, 2009.

Implementation of eLearning in Azerbaijan

Leyla Muradkhanli
Khazar University, Baku, Azerbaijan
Bahram Atabeyli
Qafqaz University, Baku, Azerbaijan

Abstract - This article discusses the experience of implementation of eLearning in Azerbaijan. Various factors contributing to the success and challenges of eLearning implementation are also discussed.

I. INTRODUCTION

The ICT sector in Azerbaijan demonstrates high rates of growth. Development of ICT in Azerbaijan has been accepted as a priority spheres in the non-oil sector. There is a great potential for future development of this sector in Azerbaijan.

The Internet in Azerbaijan is growing, supported by a national strategy to develop the country into an ICT hub for the Caucasus region [1].

Studies have shown that in 2011, 56 percent of Azerbaijan's population has at least one personal computer. On average, there are 22 computers for every 100 people. The rate of Internet penetration, according to the Azerbaijan Marketing Association, was 68 percent, or doubles the world average. According to the World Economic Forum, Azerbaijan's rating has improved by nine points and the country ranked 61 among 142 countries [2].

Azerbaijan actively integrates modern ICT into educational process. The main goal of the nation-wide initiatives is to provide and teach the use of the modern ICT tools on every level of education and thus, raise the general education standards. The Ministry of Education has been successful in implementing two State Programs of the "Provision of Information and Communication Technologies for Education" in 2005-2007 and "Informatization of Educational System" in 2008-2012.

eLearning – learning available anywhere, anytime and anyone – is one of the main growing application of ICT.

eLearning combines the internet with new multimedia technologies to provide a quality learning experience. eLearning is more than just access to resources and services. It also includes communication and collaboration between remote users. It can be used with all kinds of learners: students, employees, and employers; as well as educators who also must continue to learn.

eLearning spans a wide range of delivery methods, including web-based training, computer-based training, virtual classrooms, and digital collaboration.

During the recent years eResources successfully integrated into Azerbaijani classrooms. eLearning centers were established in many Azerbaijani universities to support and help learners.

II. THE IMPLEMENTATION OF E-LEARNING

The implementation of eLearning in any institution depends on investments of time, money, infrastructure and human resources, such as experience, skills, knowledge and attitude.

Some education institutions have already practice in eLearning by using distance education or practicing with web based courses.

The process of the implementation of eLearning involves the following steps :

- Preparation of organization for eLearning
- eLearning process
- Evaluation of eLearning
- Improvement and sustainable development

Team setup and selection of instructional model play important role in the preparation of organization for eLearning.

Team members for eLearning project are :

- Project manager
- Instructional designer
- Content developer
- Multimedia expert
- System administrator
- Trainer
- Online tutor

Selection of instructional model based on choosing one of the following model :

1. Using the technology to support the traditional face-to-face course
2. Blended learning, integrating online activities into a traditional course to enhance the learning experience
3. Delivering pure online course.

Blended learning approach that combines best practice in face-to-face and online classes is most effective and enhances and extends learning opportunities for learners [3]. Many Azerbaijani universities have adopted a blended learning approach to its learning and teaching.

eLearning process involves development and implementation of course for online delivering.

eLearning course development process is shown in Fig.1.

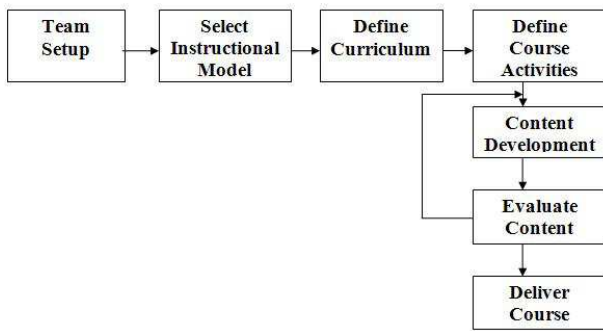


Fig. 1 eLearning course development process

The following tasks are part of the implementation and should be carried out by the project team:

- Production of content
- Provision of a learning platform
- Training of trainers and tutors for their task
- Implementation of the eLearning course
- Evaluation of the eLearning course

Implementation of eLearning depends on human, process and technology dimensions.

Human dimension

Preparation of academic staff for eLearning and training of instructors, tutors and administrators are very important. Guidelines and documentations should be prepared for this.

Process dimension includes five steps according ADDIE model (Kruse 2004).[4]:

- 1) **Analysis** - define the needs and constraints
- 2) **Design** - specify learning activities, assessment and choose methods and media
- 3) **Development** - begin production, formative evaluation, and revise
- 4) **Implementation** - put the plan into action
- 5) **Evaluation** - evaluate the plan from all levels for next implementation

The advantage of ADDIE is the independency on any particular technology. The ADDIE model stresses the important of education aspects not technological ones.

Technology plays an important role in eLearning. The main role in implementation of eLearning plays technology, which includes system selection, integration and infrastructure. In system selection organization should select Learning Management System (LMS).

LMS is a software application for the administration, documentation, tracking, and reporting of training programs, classroom and online events, e-learning programs, and training content. A robust LMS should be able to do the following [5] :

- centralize and automate administration
- use self-service and self-guided services

- assemble and deliver learning content rapidly
- consolidate training initiatives on a scalable web-based platform
- support portability and standards
- personalize content and enable knowledge reuse.
- deliver online training and webinars

This selection mostly depend on IT infrastructure (hardware, software, Internet connection and speed, security), experience and the characteristics of LMS.

Many education institutions in Azerbaijan use free and open source LMS Moodle.

The structure for implementation of eLearning is shown in Fig. 2.

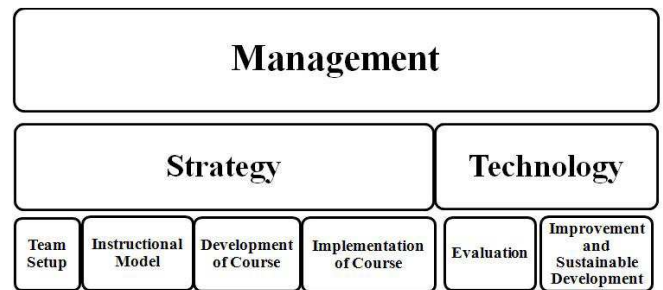


Fig.2. Implementation of eLearning

The eLearning strategy create an architecture which will promote the building of pedagogical innovation, increase the deployment of learning technologies and enable research into eLearning in a way that directly addresses business opportunities and imperatives. It provides for equivalent and enhanced learning and support all types of learner. It offers a framework that not only develops and extends the range of services and approaches already in place but also looks to deepen understanding and deployment of learning technologies.

III. BENEFITS AND CHALLENGES OF ELEARNING SYSTEM

The benefits of eLearning are

- Flexibility is a major benefit of e-learning.
- E-learning has the advantage of taking class anytime and anywhere. Learner can access to course materials 24 hours a day, and seven days a week.
- Time saving.
- E-learning is more cost effective than traditional learning because less time and money is spent travelling.
- Students have the advantage of learning at their own pace.
- eLearning is an advantage for students, especially master students who can fit e-learning into their busy job schedule.

- Student centered learning. E-learning is more focused on the learner.
- E-learning encourages students to take personal responsibility for their own learning and helps students acquire knowledge.
- Students can also learn through a variety of activities that apply to many different learning styles learners have.
- eLearning improves collaboration and interactivity among students.

eLearning implementation challenges in Azerbaijan are :

- Accessibility. Learners need to have access to a computer and Internet.
- Infrastructure and safety.
- Internet connection.
- Technical skills and support.
- University administrators support.
- Lack of expertise in instructional design and content development.
- Training of instructors, support staff and students.
- Learners' motivation. Students have to be highly motivated and responsible because all the work they do is on their own.
- Lack of financial resources.
- Intellectual property and copyright.

IV. PROMOTION OF ELEARNING IN AZERBAIJAN

GIZ "Institution Building and Human Resource Development for eLearning in the South Caucasus" program supported promotion of eLearning in Azerbaijan. The program was supported and managed by the Federal Ministry of Economic Cooperation and Development in the Federal Republic of Germany (GIZ).

Three Azerbaijani universities (Khazar University, Qafqaz University and Azerbaijan Tourism University) have built eLearning teams, developed and managed eLearning pilot projects and created crucial institutional frameworks for the implementation of eLearning. During all these steps the institutions were trained and supported by the experts of this project.

As a successful implementation of this project Azerbaijan eLearning Coordination Council was organized and Azerbaijan eLearning Network was created. In the framework of this project four new organizations – Azerbaijan University of Architecture and Construction, Nakhchivan State University, Institute of Information Technology of ANAS and Madad NGO joined to Azerbaijan eLearning Network.

Azerbaijan eLearning Network promotes the expansion and use of eLearning at universities and education and training centres in Azerbaijan.

IV.CONCLUSION

eLearning is a future for learning . E-learning is beneficial to education institutions, corporations and to all types of learners. Learners enjoy having the opportunity to learn at their own pace, on their own time, and have it less costly. E-learning is flexible and can be customized to meet the individual needs of the learners.

A successful implementation of eLearning can be achieved by:

- Management support at all levels
- Sustainable government support
- Support from university administration
- Training
- Technical support
- Policies and procedures
- e-Learning internal marketing

There is a need in Azerbaijani Universities to understand and develop effective eLearning using Azerbaijan eLearning Network experience. Azerbaijan eLearning Network will share the experience and skills in eLearning design and teaching online and support promotion eLearning in Azerbaijan.

REFERENCES

- [1] http://en.wikipedia.org/wiki/Internet_in_Azerbaijan.
- [2] Internet Penetration Rate in Azerbaijan Double World Average, trend.az, Friday, 6 April 2012.
- [3] L.G.Muradkhanli Blended Learning : The Integration of Traditional Learning and eLearning, Application of Information and Communication Technologies, AICT 2011 V International Conference, Baku, Azerbaijan, 12-14 October, 2011, pp.373-376.
- [4] Kruse, K. (2004) Introduction to Instructional Design and the ADDIE Model, eLearningGuru, [online], http://www.e-learningguru.com/articles/art2_1.htm
- [5] A field guide to Learning Management System By Ryann K. Ellis, ASTD Learning Circuits, 2009.

Modelling, simulation and monitoring the use of LabVIEW

Štefan Koprda, Milan Turčáni, Zoltán Balogh
Department of Informatics, Faculty of Natural Sciences,
Constantine the Philosopher University in Nitra
Tr. A. Hlinku 1
940 11 Nitra, Slovakia

Abstract- The article deals with the issues of modelling of loading of a computer processor using LabVIEW. In the initial part of the article the authors deal with the definition of basic terms in the area of modelling and simulation of processes. In the second part, methods of verification and validation of models are presented. The article describes the modelling environment LabVIEW, creation of virtual instruments, methods of planning and the proposal of processes, as well as distant administration of the processes. In the given environment authors designed a virtual instrument to monitor the loading of processor. The application can be used as a diagnostic tool for the tuning of processes, demanding as to calculation and memory. The application can also serve for fast processor temperature detecting. The created application can be set on in the real time in the most frequently used types of computers with up-to-date platform of processes.

I. INTRODUCTION

The main aim of modelling, the meaning of models creation, is to describe the content, structure and behavior of the real system representing a certain defined part of reality. The term model is closely connected with simulation, by means of which it is possible to represent the modelled system and its behavior also in a real time, for example on a computer, and to monitor or modify it [1], [2], [3].

A system can be defined as an object, already existing, or understood in an abstract meaning, which we intend to explore, while during its existence the system can evolve and cooperate with other systems, which form its surroundings [4]. Under the term "model" we generally understand a system, which is a certain simplification of the original of the modelled system. Between the original and its model there exists a homomorphous relationship of the display, while we differentiate between abstract models, which we can logically ponder over, and simulation models, on which we can perform simulation experiments [5], [6].

Modelling, in terms of research technology, is a replacement for the explored system by its model; its aim is to obtain information on the originally explored system by means of an experiment with the model [7].

Modelling is a multidisciplinary activity, since the knowledge of mathematics and physics, theory of systems, probability theory, informatics, cybernetics or cognitive sciences, operation research, and others, can take a share in. Modelling serves not only to solve practical problems, but it is also

designed for the realization of certain research and experiments [8], [9], or to simulate phenomena and processes.

Modelling language is every artificial language, which can be employed for the expression of information or knowledge on the systems in the structure, which is defined by a consistent set of rules. In general, they are used for the interpretation of the meaning of components in the structure. Modelling language can be graphic or textual [10].

Simulation is a method of acquisition of new knowledge on the system by means of experimenting with its model [11].

Verification of the model occurs in the moment when the model creator designedly tests seemingly correct version of the model in order to find and correct the mistakes, which could have originated during the modelling phase.

Validization occurs, when the model creator and professionals assess to what degree is the created model satisfactory and suits the original [12], [13].

It is inevitable to state that no model can be verified or validated for 100%, since neither validation nor verification is absolute. Every model is a certain representation of the system and its behaviour is even in the most ideal case only an approaching to the behaviour of the real system. If we affirm that the model was verified or validated, it means that we have carried out a sufficient number of tasks, tests and analyses. The process of verification and validation always remains to a great degree a matter of subjectivity [5].

The word "virtually" more and more frequently appears in context with new technologies connected with the fast development of information technology. We can meet virtual reality, as part of the show-business (film, computer games), or as the means facilitating coping with complex situations in certain fields of human activity. Our lives are entered by an offer of goods in virtual department stores displaying their goods by means of Internet in the so-called e-shops. Another area, connected with the word „virtual“ and the beginning of new technologies, in connection with measurement and measurement technology is called virtual setting (see Fig. 1.) [14].

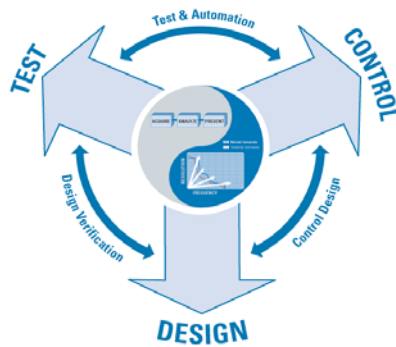


Fig. 1. Virtual setting

From among standard software tools for the area of measurement can be used for example table processors, which cover the phase of presentation and partially the one of analysis of the measured data.

According to the other point of view, in the market of software tools we can find closed systems, which provide the user with a limited variety of functions, programmed by their creator, and which cannot be extended further in a simple way.

Besides these, there exist also open systems, which provide the user with a whole range of functions, without limiting him, since they can be extended in a simple way, as to the needs of the user. These are the so-called development environments.

We shall deal with the work in the program environment LabVIEW by the firm National Instruments. LabVIEW is a development environment for the creation of applications, the so-called virtual instruments, oriented to the spheres of measurement, processing and utilization of the measured data [15].

II. LABVIEW ENVIRONMENT

A. LabVIEW

LabVIEW (Laboratory Virtual Instruments Engineering Workbench) is a development environment on the level of for example C language, but on the contrary, it is not oriented textually, but graphically. The resulting product of this development environment is called a virtual instrument, since with its outputs and activities it commemorates a classical instrument in its physical shape. The environment serves for the development of a complete system, ensuring control of the whole process of the measured data collection, their analysis and graphic presentation. Data collection can be executed from the mechanism including instruments equipped with buses GPIB, VXI, or serial interface and plug-in ISA PCI cards into personal computers. Collected data can be analysed by an extensive mathematical mechanism, including library for generating the signal, window functions, digital filters, statistics, analysis of signal in time domain and frequency domain, regression function, operations with fields and linear algebra. LabVIEW operates on platforms Windows, Linux, MacOS, SUN and HP-UX.

B. Virtual Instrument

Virtual measuring instrument (Virtual instrument - VI) is a final product and a basic unit of application of this

development environment LabVIEW. Among the advantages of virtual measuring instruments can be included lower economic costs of development and operation, open architecture, better interconnection with other facilities (e.g. relationship with networks or peripherals) in comparison with the traditional measuring instrument and faster time of variation with regard to the fact that the key component of virtual instrument is the software, which can be easily adjusted to the current requirements laid on collection, analysis and presentation of data, etc. Virtual instrument is characterized by the following features (see Fig. 2.):

- Interactive graphic interface (Graphical User Interface - GUI) – front panel, simulates the corresponding front panel of the physical instrument.
- Activity of the virtual instrument is given by its block diagram – virtual instrument has a hierarchic and modular structure.

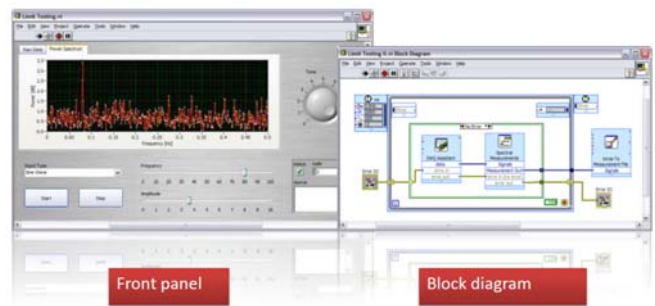


Fig. 2. Virtual instrument – front panel and block diagram

Every virtual instrument can be used as an independent application, or in the form of a subordinated virtual machine, the so-called SubVI, which can be understood by analogy as a subprogram in the text-oriented programming. Creator of the program can thus divide his application into individual parts, for which he creates the particular SubVIs. By connecting these individual SubVIs there appears the resulting VI, which can be compiled into an exe application (file *.exe), or shared library DLL, which can be used also in text-oriented program languages. These features of virtual machines ensure their hierarchic and modular structure. It should be noted that several SubVIs can be grouped into libraries (files of type *.lib).

Every VI or SubVI is formed by a pictogram with connectors, which form a set of input and output connection points. Functionality of individual SubVIs can be tested independently from others, which in connection with rich possibilities of tuning means makes the tuning of the application very simple and transparent. Tuned final compiled applications can be operated independently from the development environment, however, a support of LabVIEW Run Time Engine is inevitable.

C. Planning and Design Process

Planning and design tips for developing a LabVIEW application (see Fig. 3.).

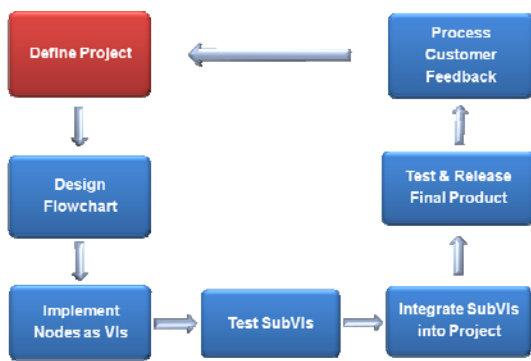


Fig. 3. Planning and Design Process in LabVIEW

This is one possible design approach, and is not intended as a general solution.

Discuss the steps in this design process:

- Clearly define project goals and system requirements.
- Design a flowchart for the application.
- Implement nodes in the flowchart as subVIs where possible.
- By creating a hierarchical set of VIs, you can find and fix bugs more quickly during testing.
- After the individual components work, begin integration into the larger project.
- Test the final product and release it.
- Use customer feedback and updated design goals to improve the product.

D. Remote control of processes

The process management task represents natural extension or completion of the process monitoring task [16]. The substantial difference between the two tasks rests in the fact that the monitoring task (access of type read only) can be utilized by more users at once on several network nodes, or on several Internet clients, while the management task (access of type write) is unique from the point of view of the process [17]. It is necessary to provide this uniqueness with an appropriate synchronization of access of individual users to the process management.

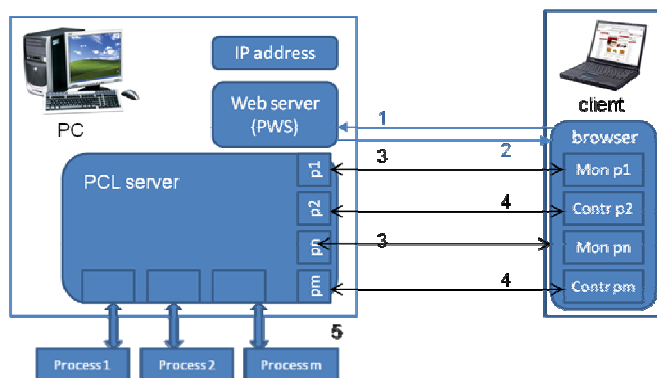


Fig.4. Object remote management principle

For the remote control of the controlled object through Internet it is possible to make use of the principle of communication between the client and the server. It is depicted in Fig. 4, where Mon p1 through Mon pn indicate application – monitoring through ports p1 and pn, Contr p2 through Contr pm indicate application – control through ports p2 through pm. In fig. 1, numeral values designate: 1 – Client’s page requirement, 2 – Server respond (page sending), 3 – communication: PCL server - applet monitor through an odd port, 4 – communication: PCL server - applet control through an even port, 5 – communication: PCL server - process through PCL card (PC LabCard). Unlike the monitoring process, where connection between the client and the server is executed through one channel (through one port), in case of control it is necessary to create a two-channel communication by means of two ports, while the other communication channel is designed for the transfer of command variable from the client to the server and through it to the process.

III. A DESIGN OF THE VIRTUAL INSTRUMENT FOR THE PROCESSOR LOADING MONITORING

In the program environment LabVIEW we created a virtual instrument, which allows us to monitor processor loading at various activated processes and the subsequent archiving of processor loading into the text file.

If we want to measure dates in real network it is important to define measurement conditions so we can collect more accurate data. Svec and Munk recommend the elimination of gratuitous network traffic and the clock synchronization between all nodes [18].

The front panel (see Fig. 5.) is formed by two displays indicating actual capacity utilization of the processor and the time flow of the processor loading history. In the bottom part of the virtual instrument is the block, which allows us to enter the file path to the file, into which we intend the data to be recorded. For the correct registration of data into the file it is inevitable to enter the file, into which the recording will be executed first, and only then we can activate the virtual instrument.

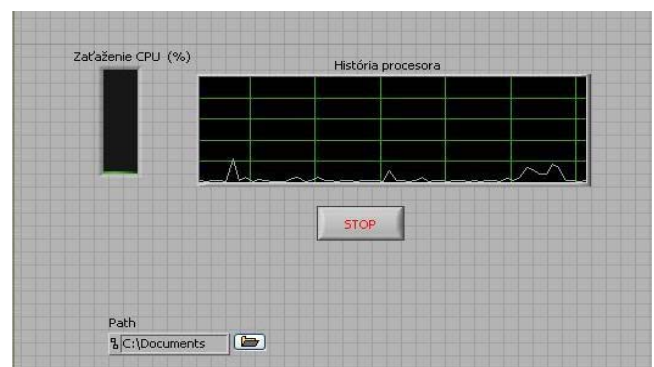


Fig.5. Virtual instrument front panel in the program environment LabVIEW

An inseparable part of the front panel (of the virtual instrument) is the block diagram, which is interconnected with the front panel, and includes individual blocks ensuring the instrument run. You can see the block diagram of the instrument for measuring and monitoring processor loading in Fig. 6.

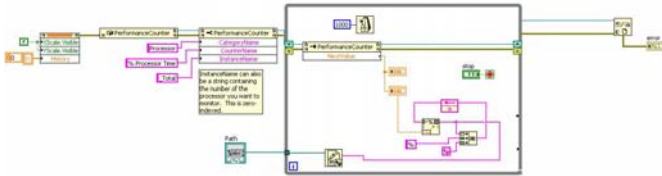


Fig. 6. Block diagram of the instrument for measuring and monitoring processor loading

For the creation of the block diagram a While loop was used. It ensures repetition of the algorithm during the validity period of the set stop condition. The cycle will be terminated only under condition that the state of the stop condition fed into the input of the condition terminal according to the previous setting will be True (Stop If True), or False (for the setting Continue If True).

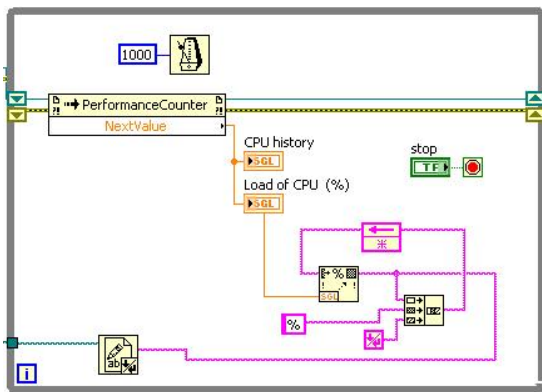


Fig. 7. Block diagram of the instrument - While loop

Inside the loop While are situated blocks, which allow us to draw processor loading into the graph. An important part of the loop is the function, which records the obtained data on processor loading in percentages into the text file. This is the Write to Text file Function. This function enters the chain or field of chains into the file as lines. In order to get the obtained data in percentages, we used in While loop the function (1), which shall allow us to enter certain inputs with values at each passing through the loop. In our case we used percentage and the following line spacing.

Outside the While loop there is the block, which is able to find out the input values on loading from the processor. Besides this, there is also the function Path, which serves for entering the file path.

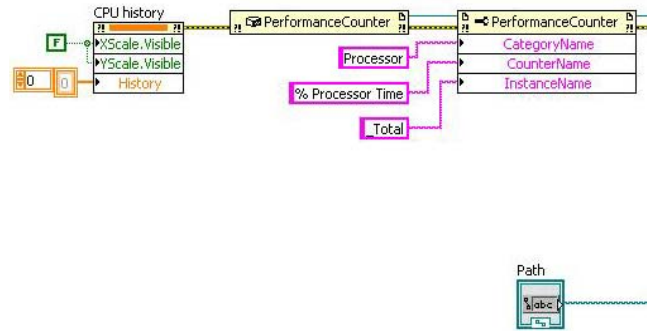


Fig. 8. Blok PerformanceCounter

Blok PerformanceCounter (see Fig. 8) is a chain allowing us to monitor a certain number of processes. In this case we used it for the monitoring of processor loading.

IV. DISCUSSION

In operating systems several processes can be under way and in the system several instances of one program can exist. The run of one process has thus an influence on the run of the other and vice versa. Operating systems, which support parallel running of several processes, normally contain synchronizing instruments, by means of which it is possible to solve synchronizing tasks.

The created application/program in the LabVIEW environment monitors the processor loading and records individual statuses of processes, which are recorded in log files, and can be subsequently evaluated by means of various statistic methods. The created application can be used as a diagnostic instrument for the tuning of processes demanding as to calculation and memory. The application can also serve for fast detection of processor temperature. By setting of appropriate levels of maximum and critical temperature, the application can protect the computer processor from the devastating effects of heat by the error tone and timely disconnection of the system.

Having finished programming of the block diagram, it is possible to remote monitor the processor loading, and in case of need, it will be possible to terminate individual processes excessively loading the processor. By means of the remote control it is possible to monitor and subsequently control the overall loading of CPU. Controlling the processes from the remote computer by means of WWW technologies provides us with numerous assets:

- Access to the controlled object from any Internet node,
- Starting this application from any platform supporting Java - browsers,
- Controlling and monitoring the processes – they can be practically verified by the use of services provided by the virtual lab. Demanding and costly installation of technological objects (or their models) is not inevitable.

V. CONCLUSION

The article presents the issues of modelling and simulation in the graphic environment LabVIEW from the firm National Instrument. The possibility to simulate real processes offers many advantages to designers and advance designers from various spheres, such as time saving and costs minimization. Prototype production is preceded, or fully replaced by the phase, in which using virtual instruments and simulation of real processes shall reveal plenty of errors leading to frequent and costly interference with the physical prototype, or total abjection of the method of solution of the given problem. Characteristic feature of the visualizing instrument - LabVIEW is utilization of the graphic environment, or G language. The designed virtual instrument is able to monitor and subsequently record processor loading in percentage. The created application can be set on various types of computers in a real time. The application was transformed to the ending exe, which means that it is not necessary to have installed the environment LabVIEW.

ACKNOWLEDGMENT

This publication is supported thanks to the Fund for supporting the Centres of Research and Development with internationally comparable quality of operations, Faculty of Natural Sciences, CPU Nitra, Slovakia.

REFERENCES

- [1] Z. Balogh, and C. Klimeš, "Modelling of education process in LMS using Petri nets structure," *Proceedings of the IADIS International Conference e-Learning 2010, Part of the IADIS Multi Conference on Computer Science and Information Systems 2010*, MCCSIS 2010, pp.289-291.
- [2] Z. Balogh, and M. Turčáni, 2011, "Possibilities of modelling web-based education using IF-THEN rules and fuzzy petri nets in LMS," *Communications in Computer and Information Science 251 CCIS (PART 1)*, pp. 93-106.
- [3] S. Koprda, Z. Balogh and M. Turčáni, "Fuzzy Control Rules Base Design," *2011 5th International Conference on Application of Information and Communication Technologies*, AICT 2011, art. no. 6110962.
- [4] A. Backlund, "The definition of system," *Kybernetes*, Vol. 29, No. 4, 2000, pp.444-451.
- [5] P. Peringer, "Modelování a simulace," *Fakulta informačních technologií: Vysoké učení technické v Brně*, 2006.
- [6] J. Fejfar, J. Šťastný and M. Cepl, "Time series classification using k-Nearest neighbours, Multilayer Perceptron and Learning Vector Quantization algorithms," *Acta Universitatis Agriculturae et Silviculturae Mendelianae Brunensis* 60 (2), 2012, pp.69-72.
- [7] I. Křivý and E. Kindler, "Simulace a modelování," *Učební texty Ostravské university, Přírodovědecká fakulta: Ostravská univerzita*, 2001
- [8] D. Klocoková, "Integration of heuristics elements in the web-based learning environment: Experimental evaluation and usage analysis," *Procedia - Social and Behavioral Sciences*, ISSN 1877-0428, 2011, vol. 15, pp.1010-1014.
- [9] D. Klocoková, and Munk, M. "Usage analysis in the web-based distance learning environment in a foreign language education: Case study," *Procedia - Social and Behavioral Sciences*, ISSN 1877-0428, 2011, vol. 15, p. 993-997.
- [10] H. Xiao, "A metamodel for the notation of graphical modeling languages," *Computer Software and Applications Conference, COMPSAC*, Vol.1, 2007, pp.219-224.
- [11] CH. Chung, "Simulation Modeling Handbook, A Practical Approach," *CRC Press*, Boca Raton, 2003.
- [12] R. G. Sargent, "Verification and Validation of Simulation Models," *Proceedings of the 2003 Winter Simulation Conference*, 2003, pp.37-48.
- [13] K. Jensen, "Coloured Petri Nets," *Basic Concepts, Analysis Methods and Practical Use*, Vol.2, Springer-Verlag, 1997.
- [14] J. Židek, "Grafické programování ve vývojovém prostředí LabVIEW," *Ostrava: Vysoká škola báňská v Ostrave*, 2002.
- [15] Z. Balogh "Modelovanie, simulácia a riadenie procesov s využitím grafického programovacieho balíka LabVIEW," *Inovácia výskumu katedier informatiky Nitra : UKF*, 2009, ISBN 978-80-8094-579-4, pp. 11-17.
- [16] P. Horovčák, and M. Rožkanin, "Metódy monitorovania technologických procesov s využitím www technológií," *Zborník referátov konferencie Automatizácia a počítače v riadení procesov*, TU Zvolen, 1998, pp.15-29.
- [17] I. Leško, D. Baluch, P. Horovčák, J. Futó, and J. Budiš, "Monitorovací systém vítacej súpravy pre účely riadenia," *Zborník referátov 9. medzinárodnej banickej konferencie, Riadenie procesov získavania a spracovania surovín*, FBBERG TU Košice 1997, pp.91-96.
- [18] P. Švec, and M. Munk, "IPv4/IPv6 performance analysis: Transport layer protocol impact to transmission time," *2011 International Conference on Internet Technology and Applications*, iTAP 2011 - Proceedings, art. no. 6006335,

Underwater Scene Characterization Using Wavelet Packet Denoising and Adaptive Contrast Stretching

Zhengmao Ye, Habib Mohamadian
College of Engineering
Southern University
Baton Rouge, Louisiana 70813, USA

Yongmao Ye
Broadcasting Department
Liaoning TV Station
ShenYang, 110004, China

Abstract – The underwater images are always subject to severe quality degradation, which arises from sensor nonlinearity, water dispersion, wave turbulence, atmospheric fluctuation and relative motion. These factors will lead to the distortion of digital images with the limited dynamic range being further reduced. Advanced image processing techniques should be introduced for denoising, tuning and restoration of actual scenes. It is focused on producing and reestablishing an array of pixels for object representation to enhance the source images from direct measurements. Image denoising aims to remove various noises and retain important feature as much as possible. Nonlinear wavelet transform is a feasible approach to filter out the blurry underwater images. At each level of wavelet decompositions, an image is split into four subbands, representing approximation (low frequency feature) and three details (high frequency features) in horizontal, vertical and diagonal directions. For multiple level decompositions, the approximation will always be decomposed at each level no matter if discrete wavelet transform or wavelet packet transform is used. If detail components are further decomposed similar to that of the approximation, it is wavelet packet transform. If not, it turns out to be discrete wavelet transform instead. Soft thresholding at multiple levels of wavelet decomposition is an efficient method for reduction of noises. Color balance can then be applied to adjust intensities of the RGB colors (red, green, and blue), respectively, similar to the gray level intensity. Underwater color changes depend on wavelength of light, which is corresponding to diverse levels of attenuation, rendering ambient illumination dominated by a blue tone but lack of a red tone. The color balance will change the overall mixture of color components in an image after render three specific true colors to make better visual appealing. Adaptive contrast stretching is proposed for image enhancement based on generalization of histogram equalization. Dramatically diverse outcomes are reached with relatively less modifications, where the accumulation function is used to generate a grey level mapping from local histograms. By tuning the parameters, the process will produce adaptive equalization with better degrees of contrast enhancement.

Keywords — *Wavelet Packet, Image Denoising, Adaptive Contrast Stretching, Adaptive Histogram Equalization, Color Balance*

I. INTRODUCTION

In digital image processing, the energy distribution of a light source depends not only on image coordinate (x, y, z) but also on the time and wavelength. The image matrix function values correspond to either brightness or darkness at each pixel. The underwater imaging is affected by various physical quantities such as atmospheric dispersion, variations in temperature and pressure, changes in illumination, target motion, and so on.

The primary brightness integrates various optical quantities simultaneously with involvement of diverse types of noises, either slowly varying or rapidly varying. In fact, Images acquired by modern sensors are inevitably contaminated by a variety of noises (such as thermal noise, amplifier noise, photon noise, quantization noise and cross talk), resulting from deterministic distortions or shading and stochastic variation. Linear denoising consists of triangular filter, uniform filter and Gaussian filter, while median filter and average denoising are typical examples of nonlinear denoising. However, median filter could be insensitive to noise spikes and it preserves blurring edges with the iterative application. Effective image denoising is to maintain the energy of target images and reduce the energy of noises, thus wavelet transform can be introduced. After splitting an image using a set of wavelet bases, wavelet coefficients could be thresholded to reduce effects from noises. Unlike linear denoising which results in significant information loss, wavelet denoising is able to eliminate various noises and preserve major characteristics by removing high frequency components and retaining the lower ones. It involves three steps of wavelet transform, nonlinear thresholding and inverse wavelet transform. For discrete wavelet transform at the multiple levels, only approximation will be further decomposed. The information loss between two successive approximations is reflected by detail coefficients. In wavelet packet transform, details are decomposed similarly across the approximation splitting. A complete quaternary tree is produced in 2D wavelet packet transformation case [1-10].

The fractal-based denoising in the wavelet domain has been used to predict the fractal code of a noiseless image from its noisy observation. The cycle spinning is incorporated into the fractal-based denoising methods to produce better estimations of the denoised images. Application of the wavelet packet filter bank for image denoising is proposed. A data-dependent orthogonal analysis filter bank is constructed so as to obtain the signal energy compaction with some dominant principal components, while the noise is spread over all transformed coefficients. The noise is removed without blurring edges and important image characteristics [11-12]. The wavelet packet decomposing is an expansion of the wavelet decomposing. For rich texture Synthetic Aperture Radar images, high-frequency coefficients from wavelet decomposing contain abundant texture information. It decomposes both low-frequency and high-frequency coefficients. Wavelet packet denoising is more

efficient in removing noises and keeping detail information [13]. The approach has been proposed, constituting new image de-noising based on Wiener filtering for soft thresholding. It shows a high and stable SNR gain for all noise models being tested. This process improves the phase image when the real and imaginary parts of the wavelet packets coefficients could be filtered independently [14].

No existing underwater processing techniques can handle both light scattering and color change distortions for the underwater images, with the possible presence of artificial lighting simultaneously. Light scattering and color change are two major sources of distortion for underwater photography. Light scattering is caused by light incident on objects that are reflected and deflected multiple times by particles present in the water before reaching the camera. This in turn lowers the visibility and contrast of the image captured. A dehazing algorithm has been proposed to compensate for the attenuation discrepancy along propagation. Self-tuning image restoration filtering is presented which is ideally suitable for shallow-water and diffuse-light conditions with limited backscatter. Optimal values of the filter parameters are estimated automatically for each individual image by optimizing quality criterion based on a global contrast measure. The simplified model is ideally suitable for diffuse-light imaging with limited backscatter, but qualitative tests show good performance in various imaging conditions [15-16]. There are various applications of adaptive image contrast limited enhancement based on generalization of the histogram equalization. By choosing alternative forms of accumulation functions, one can achieve a wide variety of effects, which is used to generate a grey level mapping from the local histogram. Through the variation of one or two parameters, the resulting process can produce a range of degrees of contrast enhancement, at an extreme yielding full adaptive equalization [17]. In this article, fundamental image processing is conducted to render specific colors and achieve correct neutral color balance for image enhancement under shallow water, deep water and ocean floor. Nonlinear wavelet packet denoising is applied at first to extract intrinsic information. Adaptive histogram equalization is then implemented for contrast stretching, producing sharper images with better color fidelity and visual appealing.

II. WAVELET PACKET DENOISING

In order to extract intrinsic information from underwater images, it is necessary to design preprocessing techniques to compensate for the blurring and noisy source images so as to achieve high Signal-to-Noise Ratios (SNRs). It will lead to high quality results which exhibit better feature patterns upon restoration for further analysis. The noise artifacts have the negative impact on the visibility quality and interpretation of underwater images to varying degrees. However, properties and mechanisms of different types of noises vary a lot. Noises could be classified into either slowly varying noise or rapidly varying noise. For slowly varying noise, a linear filter could be enough for noise denoising. For rapidly varying noise, instead, nonlinear filter must be applied. Denoising is to suppress

noises or small high frequency fluctuations on a digital image. Wavelet packet decomposition has been proposed.

A. Wavelet Packet Transform

Discrete wavelet decomposition has been implemented as a multiple level transformation. Outputs at each level contain approximation, horizontal detail, vertical detail and diagonal detail. The quaternary tree is produced in the two dimensional case. Each has a quarter size of the source image followed by downsampling by a factor of two. For multiple-level discrete wavelet packet decomposition, both approximation and detail components are further decomposed similar to that of the first level. On contrast, for discrete wavelet transform at multiple levels, only approximation is further decomposed and detail components are not. This generates the priority of wavelet packet decomposition over discrete wavelet transform with less information loss at a tradeoff of extra computation cost. In fact, wavelet packet decomposition produces a large number of bases. In wavelet packet decomposition, digital images will pass through more filters than the discrete wavelet transform, so that the detail and approximation coefficients are subject to where both high and low pass filters. The information loss between two immediate approximations is exhibited as detail coefficients. At each level, one relatively smoother component and 3 relatively coarser components are all selected.

Discrete wavelet decomposition is formulated as below. The wavelet transform uses a set of basis functions. In a two dimensional case, a scaling function $\phi(x, y)$ and three wavelet functions $\psi^H(x, y)$, $\psi^V(x, y)$ and $\psi^D(x, y)$ are derived. Each scaling function or wavelet function is the product of the basis wavelet functions. The products produce the scaling function (1) and separable directional sensitive wavelet functions (2)-(4), resulting in structure of the quaternary tree. As an example of Haar Transform, the scaling function and wavelet functions are determined.

$$\phi(x, y) = \phi(x)\phi(y) \quad (1)$$

$$\psi^H(x, y) = \phi(y)\psi(x) \quad (2)$$

$$\psi^V(x, y) = \phi(x)\psi(y) \quad (3)$$

$$\psi^D(x, y) = \psi(x)\psi(y) \quad (4)$$

Each wavelet measures variations of digital images along three directions, where $\psi^H(x, y)$, $\psi^V(x, y)$ and $\psi^D(x, y)$ measure directional variations along directions of columns (horizontal), rows (vertical) and diagonals (diagonal), respectively. The scaled and translated basis functions are defined by:

$$\Phi_{i,m,n}(x, y) = 2^{j/2} \phi(2^j x - m, 2^j y - n) \quad (5)$$

$$\Psi^i_{i,m,n}(x, y) = 2^{j/2} \psi^i(2^j x - m, 2^j y - n), i = \{H, V, D\} \quad (6)$$

where index i identifies the directional wavelets of H, V, and D. The discrete wavelet decomposition of function $f(x, y)$ of size M by N is formulated as:

$$w_\phi(j_0, m, n) = \frac{1}{\sqrt{MN}} \sum_{x=0}^{M-1} \sum_{y=0}^{N-1} f(x, y) \phi_{j_0, m, n}(x, y) \quad (7)$$

$$w_\psi^i(j, m, n) = \frac{1}{\sqrt{MN}} \sum_{x=0}^{M-1} \sum_{y=0}^{N-1} f(x, y) \psi^i_{j, m, n}(x, y) \quad (8)$$

where $i=\{H, V, D\}$, j_0 is the initial scale, $w_i(j_0, m, n)$ coefficients define the approximation of $f(x, y)$, $w_\psi^i(j, m, n)$ coefficients represent the horizontal, vertical and diagonal details for scales $j \geq j_0$. Here $j_0 = 0$ and select $N + M = 2^j$ so that $j=0, 1, 2, \dots, J-1$ and $m, n = 0, 1, 2, \dots, 2^j - 1$. $f(x, y)$ can be further restored via the inverse discrete wavelet transform.

B. Thresholding

For wavelet packet denoising, wavelet packet coefficients at each level are subject to thresholding. In order to reconstruct an estimated image based on wavelet coefficients, the features of interest must be well preserved ahead of the thresholding. The chosen shrinkage function of soft thresholding on the wavelet coefficients is used to remove high frequency artifacts without distorting intrinsic features of source images. The approximation and detail components at both levels one and two will be thresholded for pending image reconstruction. In a thresholding process, threshold selection is upmost important. Both the mean and median values can be used as the threshold. In current study, soft thresholding is selected rather than hard thresholding, which shrinks the nonzero wavelet coefficients towards zero. In general, low thresholds produce good but still noisy estimations and high thresholds produce smooth but blurring estimations, a median value is chosen accordingly, which arises from the absolute value of wavelet coefficients at each decomposition level. The shrinkage function for soft thresholding is formulated as (9):

$$f(x) = \text{sgn}(x)(|x| - TH) \tag{9}$$

where TH is the median threshold value based on wavelet coefficients, x is the input signal, $\text{sgn}(x)$ is the sign of x and $f(x)$ is the output signal after thresholding.

After wavelet packet decomposition and thresholding, the denoised image $f(x, y)$ via wavelet packet transform can be reconstructed via inverse discrete wavelet transform as (10).

$$f(x,y) = \frac{1}{\sqrt{MN}} \sum_m \sum_n w_\phi(j_0, m, n) \phi_{j_0, m, n}(x, y) + \frac{1}{\sqrt{MN}} \sum_{i=H, V, D} \sum_{j=j_0}^{\infty} \sum_m \sum_n w_\psi^i(j_0, m, n) \psi_{j_0, m, n}^i(x, y) \tag{10}$$

Three different cases are taken into account, ranging from shallow water (15 feet), deep water (125 feet), to the ocean floor (150 feet) Using wavelet packet transform for image denoising, the filtered digital underwater images are generated which are shown in Fig. 1 to Fig. 3. Relatively sharper images are observed in each individual case. The role of wavelet packet denoising varies little across cases under shallow water, deep water and ocean floor environment.

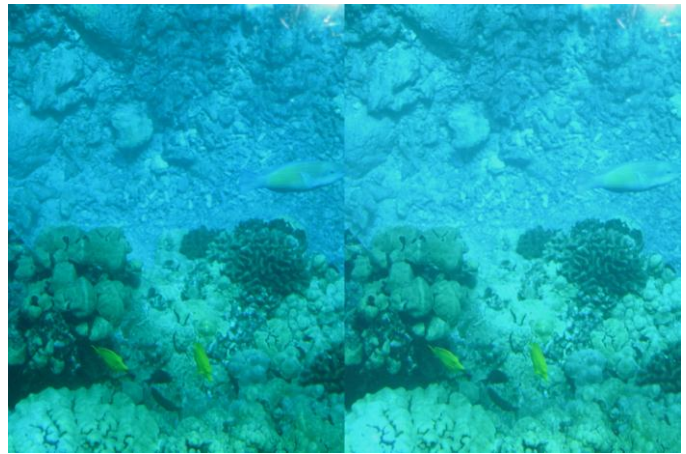


Fig. 1 Wavelet Packet Denoising of Shallow Water Image



Fig. 2 Wavelet Packet Denoising of Deep Water Image



Fig. 3 Wavelet Packet Denoising of Ocean Floor Image

III. ENHANCEMENT VIA ADAPTIVE CONTRAST STRETCHING

A. True Color RGB Model

In the true color system, each color represents a primary spectral component (Red – Green - Blue) in the Cartesian coordinate system. Each of the three intensity components of the true color system could be computed and analyzed independently before the composition true color is reached. In true color subspace, each color is uniquely mapped into a cube in which RGB values are at three corners; black locates at the origin and white locates at the corner opposite to the origin; while other three colors (cyan, magenta, yellow) locate at three remaining corners. Generally speaking, the lowest value corresponds to color of black and the highest to color of white. The grey scale lies along with the diagonal line that links black and white points. Each color acts as a vector on or inside the cube pointed from the origin. Image quantization is the A/D transition between continuous image values and its digital integer equivalent. Certain component amounts of red, green and blue will generate any particular true color. The intensity component is the composite color image from three primary image planes.

B. Histogram Contrast Stretching

Contrast enhancement produces better perception against the limitation of sensing when low contrast images are captured under poor illumination with the limited dynamic range during image acquisition. The goal of contrast stretching is to enlarge the dynamic range of digital images where sharpening is one of standard techniques. It improves the contrast within a local region instead of a broad region. The larger the contrast is, the more the enhancement is. For underwater images, contrast stretching is an effective approach. The local method covers a small neighborhood of a pixel from an input image to produce the brightness value for the output image. In adaptive neighborhood preprocessing, the neighborhood size and shape are dependent on image attributes and parameters that reflect homogeneity measure. Adaptive neighborhood is constructed for each seed pixel individually, which employs uniform distribution for contrast transformation creation. All connected pixels involve in contrast transformation. It is feasible to enhance contrast at regions rather than at borders exclusively, so that objects of interests are associated with the covering neighborhood. Contrast stretching is essentially nonlinear which can be mapped into a specified modification curve.

The high contrast level and low contrast level will display distinguishable degree of variations in true color or gray level visual perception. Highlights and shadows will depict intense differences of density in image tones for high contrast images. However low contrast images are corresponding to smaller differences of density in image tones. Root mean square (RMS) contrast is used which is defined as the standard deviation of the pixel intensities. It is not dependent on the spatial distribution of the image contrast.

$$\text{Contrast} = \frac{\sum_{i=0}^{M-1} \sum_{j=0}^{N-1} [g(i,j) - g_{AVG}]^2}{M*N} \quad (11)$$

where an intensity $g(i, j)$ denotes an element at coordinates i and j on a 2D image of the size $M \times N$. g_{AVG} is the average intensity of all pixel values within an image.

In contrast stretching, foreground and background adaptive neighborhoods are designed, where the background size is comparable to the foreground size. A contrast-limited adaptive histogram operates on small regions rather than the entire digital image. Within each region, interpolation is needed to calculate brightness of those pixels not measured. Bilinear interpolation is applied which has the advantage of utilizing small decrement in resolution, so as to eliminate artificially induced boundaries. Exponential distribution is selected to create contrast transform functions for underwater images. The constraint is set up for the contrast inside homogeneous areas to avoid amplifying noises. Contrast stretching is conducted adaptively around the selected seed pixel's neighborhood until the satisfactory contrast is achieved.

$$\text{Out}(i, j) = M \frac{\text{In}(i, j) - \text{Min}(i, j)}{\text{Max}(i, j) - \text{Min}(i, j)} \quad (12)$$

where $\text{In}(i, j)$ is brightness value of measured seed pixel, M is the contrast limitation factor that is chosen to produce the desired dynamic range and to prevent over-saturation of the image in homogeneous areas. Max and Min are maximal and

minimal values around the seed pixel neighborhood, $\text{Out}(i, j)$ is the enhanced brightness value after processing. The optimal number of small data region depends on input image itself.

C. Adaptive Contrast Stretching

Adaptive contrast enhancement schemes based on the histogram equalization are applied. Similar to the processing of gray level images, each of three primary color components is processed individually by adaptive histogram equalization algorithms. Each component of a true color image is split into a number of small regions. Within a small region, histogram is calculated with the contrast constraint, exponential distribution is then applied as a basis to create the contrast transform function. The assignment of pixels in each small local region is specified by contrast transformation, so that mapping from local histogram is generated. To avoid the occurrence of boundary artifacts stem from neighborhoods of small regions so as to generate an evenly distributed smooth enhancement across entire regions, bilinear interpolations are applied and contrast saturation is solved using constraints, until the final contrast from adaptive histogram equalization is satisfactory. Excessive enhancement of noises is avoided when using the adaptive schemes. It has been applied to case studies of a set of underwater true color images to be characterized, where the histogram of each color component (Red, Green and Blue) contains 256 bins. The percentage of counts for each bin over the total accumulation value will produce the probability distribution. Tuning will be made for better quality on the parameter of exponential functions, factor of contrast limiting and weight of bilinear interpolation.

IV. CASE STUDIES

For underwater images, color changes depend on amount of attenuation associated with the wavelength of each color component. The ratios of the residual energy on three different color channels depend on the water depth due to discrepancy in wavelength attenuation in the direction of the water wave propagation. Underwater scenes being captured are essentially dominated by the blue tone but suffered from the lack of red tone. Compensation is necessary to preserve the color fidelity via appropriate tuning of color balance, for examples, to strengthen the red tone and to weaken the blue tone. In this case, suitable color balance with sharper visibility and less haze can be observed. By adaptive contrast stretching of 3 primary color components followed by fine tuning of color balance, some preliminary results have been obtained. From Fig. 4 to Fig. 6, three case studies are conducted representing denoised source images and restored images using adaptive contrast stretching and fine tuning. It is shown that all the reconstructed images from shallow water, deep water and ocean floor provide sharper images with better color fidelity.

A. Enhancement of Shallow Water Image

For shallow water images, the loss in red tone could be easily compensated and the extra gain in blue tone could be easily reduced. The color change in green tone is also subject to minor adjustment. As a result, remarkably enhanced true color underwater images are observed by means of the

proposed integration of the wavelet packet denoising, adaptive contrast stretching and color balance tuning schemes, which is shown in Fig. 4. The enhanced true color image is shown, followed by its red, green, and blue components.

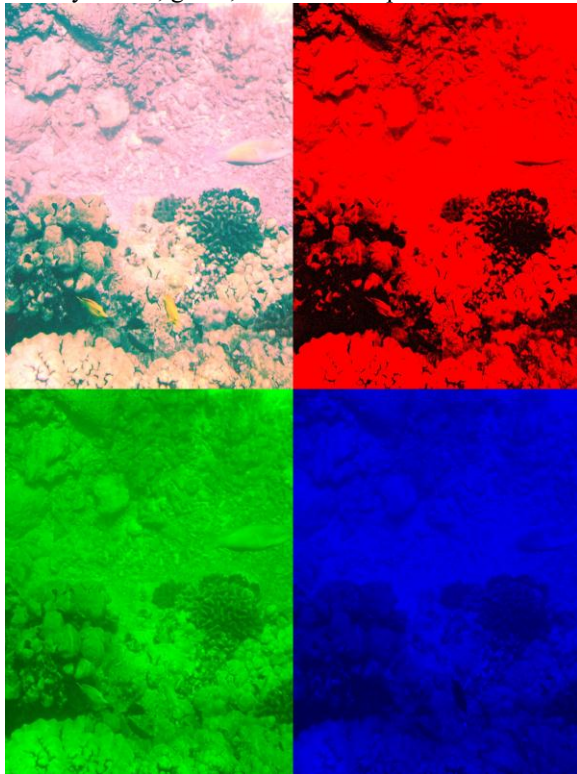


Fig. 4 Contrast Stretching and Color Tuning of Shallow Water Image

B. Enhancement of Deep Water and Ocean Floor Image



Fig. 5 Contrast Stretching and Color Tuning of Deep Water Image

For deep water and ocean floor images, it is relatively tough to compensate for the loss in red tone and extra gain in blue done. However, enhanced true color underwater images can still be obtained via technology integration, where extra efforts in color balance tuning are necessary. The results of two case studies are shown in Fig. 5 and Fig. 6, respectively. The enhanced true color images are shown, followed by the red, green, and blue components. From the available simulation

results, quality enhancement of the deep water image is higher than that of the ocean floor image, but lower than that of the shallow water image.

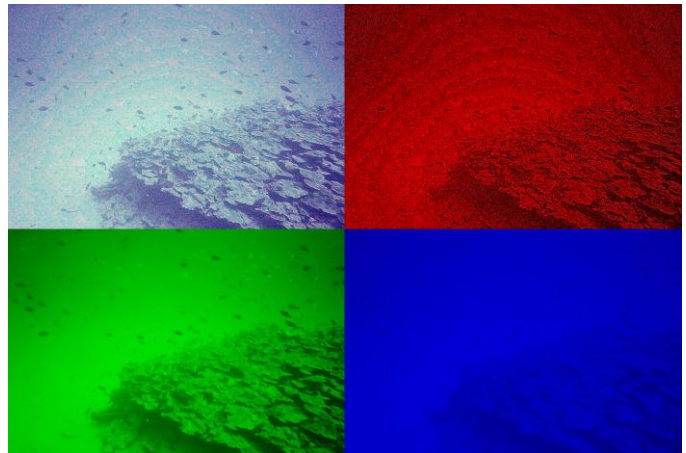


Fig. 6 Contrast Stretching and Color Tuning of Ocean Floor Image

From Figs. 4-6, it has been illustrated that for all three cases, dramatic views with the broader dynamic range are generated.

CONCLUSIONS

This research is concerned with the scene characterization of underwater image patterns in terms of water dispersion and atmospheric variation. Noise filtering is an important step to extract the intrinsic information in image processing, thus the wavelet packet approach is employed with satisfied denoising results. Considering 3 cases of shallow water, deep water and ocean floor, underwater environment is affected at varying levels. In most cases the underwater images are dominated by a blue tone and suffered from lack of a red tone, however. Hence the major objective is to render specific color mixture based on color adjustment of individual components of red, green and blue, adaptive tuning based histogram equalization and contrast stretching approaches are proposed to correct the color balance of red, green and blue. Based on outcomes of three case studies, integration of wavelet packet denoising and adaptive contrast stretching always gives rise to high dynamic range scenes with sharper patterns. Significantly better visual appearance is achieved after image enhancement and color correction. At the same time, the quality of color correction depends on the actual water depth. Better color correction is made from the shallow water scene.

REFERENCES

- [1] R. Gonzalez and R. Woods, "Digital Image Processing," 2nd Edition, Prentice-Hall, 2002
- [2] A. Engelbrecht, "Computational Intelligence: An Introduction", 2nd Edition, John Wiley & Sons, 2000
- [3] Simon Haykin, "Neural Networks - Comprehensive Foundation", 2nd Edition, Prentice Hall, 1999
- [4] A. Shapiro, G. Stockman, C. George, "Computer Vision". Prentice Hall. ISBN 0-13-030796-3, 2002

- [5] Z. Ye, H. Cao, S. Iyengar and H. Mohamadian, "Medical and Biometric System Identification for Pattern Recognition and Data Fusion with Quantitative Measuring", Systems Engineering Approach to Medical Automation, Chapter Six, Artech House Publishers, pp. 91-112, ISBN978-1-59693-164-0, October, 2008
- [6] Z. Ye, H. Mohamadian and Y. Ye, "Information Measures for Biometric Identification via 2D Discrete Wavelet Transform", Proceedings of the 2007 IEEE International Conference on Automation Science and Engineering, pp. 835-840, Sept. 22-25, 2007, Scottsdale, Arizona, USA
- [7] Z. Ye, H. Mohamadian and Y. Ye, "Quantitative Effects of Discrete Wavelet Transforms and Wavelet Packets on Aerial Digital Image Denoising", Proceedings of the 2009 International Conference on Electrical Engineering, Computing Science and Automatic Control (CCE 2009), pp. 416-420, November 10-13, 2009, Toluca, Mexico
- [8] Z. Ye, "Objective Assessment of Nonlinear Segmentation Approaches to Gray Level Underwater Images", International Journal on Graphics, Vision and Image Processing, ISSN 1687-398X, pp. 39-45, Volume 9, Issue II, April, 2009
- [9] Z. Ye, Y. Ye and H. Mohamadian, "Biometric Identification via PCA and ICA Based Pattern Recognition", Proceedings of the 2007 IEEE International Conference on Control and Automation (ICCA 2007), pp. 1600-1604, May 30-June 1, 2007, Guangzhou, China
- [10] Z. Ye and G. Auner, "Linear Filtering and Nonlinear Fuzzy Logic Filtering for Sample Identification with Raman Spectroscopy", Proceedings of the 2003 IEEE International Conference on Systems, Man and Cybernetics (SMC 2003), pp. 4619-24, Oct 5-8, 2003, Washington, DC, USA
- [11] M. Ghazel, G. Freeman, and E. Vrscay, "Fractal-Wavelet Image Denoising Revisited", IEEE Transactions on Image Processing, Vol. 15, No. 9, September, 2006
- [12] S. Bacchelli and S. Papi, "Image Denoising Using Principal Component Analysis in the Wavelet Domain", Journal of Computational and Applied Mathematics 189, pp. 606-21, 2006
- [13] W. Wang, X. Yi and P. Fei, "Denoising of SAR Images Based on Lifting Scheme Wavelet Packet Transform", Geospatial Information Science 11(4): 257-261, Volume 11, Issue 4, 2008
- [14] J. Lorenzo-Ginori and H. Cruz-Enriquez, "De-Noising Method in the Wavelet Packets Domain for Phase Images", CIARP 2005, LNCS 3773, Springer-Verlag, pp. 593 – 600, 2005
- [15] E. Trucco and A. T. Olmos-Antillon, "Self-tuning underwater image restoration," IEEE J. Ocean. Eng., vol. 31, no. 2, pp. 511-519, Apr. 2006
- [16] J. Chiang and Y. Chen, "Underwater Image Enhancement by Wavelength Compensation and Dehazing", pp. 1756-1769 IEEE TRANSACTIONS ON IMAGE PROCESSING, VOL. 21, NO. 4, APRIL 2012
- [17] K. Zuiderveld, "Contrast Limited Adaptive Histogram Equalization", Graphics Gems IV, AP Professional, 1994, pp. 474-485
- [18] J. Stark, "Adaptive Image Contrast Enhancement Using Generalizations of Histogram Equalization", IEEE TRANSACTIONS ON IMAGE PROCESSING, pp. 889-896, VOL. 9, NO. 5, MAY 2000

Computational Study of the Conformational Flexibility of the Amphibian Tachykinin Neuropeptides

G.A.Agaeva

Institute for Physical Problems, Baku State University,
AZ-1148, Baku, Z.Khalilov Str.23, Azerbaijan
gulshen@mail.ru

Abstract- The conformational flexibility of some amphibian tachykinin neuropeptides have been investigated by computer modeling with molecular dynamics method in the different conditions. At the first stage the conformational changes of these peptides were studied in vacuum, but in the second stage they were surrounded by water molecules with the periodic boundary conditions. All molecules were observed in vacuum and in water with large flexibility of the N-terminal parts of its amino acid sequences. It is shown that C-terminal backbone parts of these molecules save a alpha-helix conformation, but their side chains may exist in more than one orientations in all conditions.

Keywords: amphibian tachykinin, structure, function, conformation, neuropeptide

I. INTRODUCTION

Tachykinins are a family of biologically active peptides distributed in the central and peripheral nervous system. The earliest known members of the tachykinin family are those that are present in mammalian systems. Tachykinins elicit a wide and complex array of biological responses, such as the stimulation of extravascular smooth muscle, powerful vasodilation, hypertensive action, activation of immune system, regulation of pain transmission, and neurogenic inflammation. The tachykinin peptides are characterized by a common C-terminal sequence, Phe-X-Gly-Leu-Met-NH₂, where X represents either an aromatic (Phe, Tyr) or a branched aliphatic (Val, Ile) amino acid. The C-terminal region or the message domain is considered to be responsible for activating the receptor. The divergent N-terminal region or the address domain varies in amino acid sequence and length and is postulated to play a role in determining the receptor subtype specificity. Three pharmacologically distinct receptor subtypes have been identified and cloned for tachykinins designated as NK-1, NK-2, and NK-3, which all share a significant sequence similarity. The wide range of physiological activity of tachykinins has been attributed to the lack of specificity of tachykinins for a particular receptor type. This lack of specificity can be accounted for by the conformational flexibility of these short, linear peptides. The characterization

of the biologically active conformation, which controls receptor binding and subtype selectivity, is of significant interest. The most studied tachykinin, substance P (SP), was first isolated from equine brain and intestine. Its amino acid sequence was identified as Arg1-Pro2-Lys3-Pro4-Gln5-Gln6-Phe7-Phe8-Gly9-Leu10-Met11NH₂. SP is involved in transmission of pain, causes rapid contractions of the gastrointestinal smooth muscle, and modulates inflammatory and immune responses. In spite of extensive structural research of the SP and its analogs by different spectroscopic methods, is very difficult to determine of selective antagonists of tachykinin receptors. Physalaemin is a undecapeptide with the sequence pGlu1-Ala2-Asp3-Pro4-Asn5-Lys6-Phe7-Tyr8-Gly9-Leu10-Met11NH₂. It was first isolated from the *Physalaemin fuscumaculatus* amphibian skin. Uperolein belongs to the tachykinin family which excites neurons, provoke behavioral responses, are potent vasodilators and secretagogues. Uperolein is an 11-peptide with the sequence pGlu1-Pro2-Asp3-Pro4-Asn5-Ala6-Phe7-Tyr8-Gy9-Leu10-Met11-NH₂ and it isolated from the skin of the Australian frog *Uperoleia rugosa* belonging to the tachykinin family neuropeptides. Uperolein, a physalaemin-like endecapeptide, has been shown to be selective for Neurokinin 1 receptor. It is known that SP, physalaemin and uperolein interact with one subtype of NK receptors, therefore physalaemin in the literature is often called as a structural analog of substance P [1-7]. For understanding of how ligand interact with their receptor is required the knowledge of the conformational specificity and dynamics of the native molecule allowing a rational design of compounds acting selectively at the tachykinin receptor level. The major aim of the present article is the investigation of the conformational dynamics for SP, physalaemin and uperolein, with the purpose of getting insight into basic structural requirements that determine ligand-receptor interaction. The conformational particularities of SP, physalaemin and uperolein conformational dynamics of its backbone and side chains at the present article have been investigated by molecular dynamics methods in vacuum and water molecules surrounding. In general, it has been found that

the tachykinins display some elements of secondary structure in appropriate solution environment, though it has been suggested that they do undergo rapid conformational exchange. There are no discernible trends in the conformation of the address segments of these peptides. However, the message domains are similar in each case. In general, the message domain of these peptides undergoes conformational averaging in aqueous environments. In hydrophobic environments, the message domain assumes helical conformations or exists as a series of turns in dynamic equilibrium. It has been postulated that the binding of neuropeptides to their cell surface receptors may be catalyzed by nonspecific interactions with membrane lipids and the binding of the peptide to the receptor occurs in at least two sequential steps: the binding of the peptide to the membrane, followed by the binding of the peptide to the receptor in the membrane. Although the neuropeptides in aqueous solution exist as randomly distributed conformers, the biologically active forms of these neuropeptides are likely to be ordered and stabilized within the lipid bilayers of the cell membrane before binding with their receptors. Some types of stable conformations with significantly different values of dihedral angles are determined for three tachykinins. A molecular conformation is largely determined by its environment, so the aim of this present work is the study the differences in the conformations of the tachykinin peptides in a vacuum and in aqueous environment using a molecular dynamics method.

II. COMPUTATIONAL METHOD

To understand the physical and structural properties of membrane-bound peptides and proteins and their relationship to the biological activities, Molecular dynamics (MD) simulations with everimproving force fields and longer time scales have been providing molecular level details of such systems. MD simulations were performed for neuropeptides in vacuum as well as in water solution using modeling package [8]. MD is widely applied to the study of biological systems, providing insight into the structure, function, and dynamics of biological molecules [9,10]. A wide range systems have been treated, from small molecules to proteins, in vacuum and in the presence of solvent [11, 12]. Molecular dynamics simulations generate trajectories of atomic positions and velocities and some general thermodynamic properties. MD involves the calculation of solutions to Newton's equations of motions. Often an MD trajectory will become trapped in a local minimum and will not be able to step over high energy conformational barriers. Thus, the quality of the results from a standard MD simulation is extremely dependent on the starting conformation of the molecule. So, the low-energy structures, including the best and the worst of the calculated structures from [6] and [7] were used as starting conformations for molecular dynamics simulations ϕ , ψ and χ angles were analyzed for changes in each conformation. Runs were performed for 300 ps at 300K. The total length of the simulation depend on the system being studied and the type of information to be extracted. For example, in simulations of

biological system a time step of 1 fem to second is commonly used. To ensure that information about the highest frequency in the system is $\pm 30^\circ$ about a mean position during the molecular dynamics simulations. The length of the simulation (after equilibration) has to be long enough to enable the slowest modes of motion to occur.

The force field parameters were those of the all atom version of AMBER by Cornell et al [18]. A harmonic force towards the center of the sphere was added to atoms when they moved out of the sphere. The nonbonded cutoff distance was 12Å. The time step was 0,5fs. The program Hyper. Chem. 7.01 [19] was used for the MD simulations.

III. RESULTS AND DISCUSSION

Preliminary theoretical conformational analysis of SP has shown that its spatial structure may be described by five families of low-energy conformations with identical structure of the C-terminal heptapeptide. The C-terminal part (residues 5-11) of the SP can adopt a partially helical structures, but the N-terminal part is different in each family. Only four low-energy conformations of the SP are fall in the 0-5 kcal/mol energy interval. It is shown that two preferred conformations have very similar backbone form and values of the relative energy. In these conformations only Gln⁵ residue is in the different backbone forms. Both conformations contain some turn at the N-terminal tripeptide, but the first conformation have β -turn structure at the Pro⁴-Gln⁵-Gln⁶-Phe⁷ segment also. These β -turns are confirmed by distance between C ^{α} atoms of the i and i+3 residues ($< 7 \text{ \AA}$). The lowest energy structures of SP exhibit the most favourable dispersion contacts and therefore may be expected to become the most preferred in a strongly polar medium, when electrostatic interactions do not play a significant role. The lowest energy α -helical structure at the C-terminal fragment is stabilized by network of hydrogen bonds. Theoretical conformational analysis of physalaemin have been indicated four families of low-energy conformations with similar C-terminal heptapeptides. Unlike the molecule of substance P, this analysis has shown that physalaemin can form one global, i.e. the lowest-energy structure, which is consist one β -turn on Asp³-Pro⁴-Asn⁵-Lys⁶ segment and on the C-terminal part α -helical segment, formed by regular hydrogen bonds.

MD simulation, using the four starting lowest energy structures of each molecule from [6] and [7] were shown the significant differences in the conformations of the molecule in a vacuum and in an aqueous environment. Structural reorganization of the global conformation of all peptides at the molecular dynamic simulation in vacuum and in water solution are obtained. Figure 1 shows the global structures of the molecules as a result of the molecular simulation of aqueous environment. The MD simulations revealed the possible deviation by $\pm 10^\circ$ from the optimal values of ϕ , ψ , ω , χ dihedral angles in vacuum as compared to $\pm 20^\circ$ in water. The permissible changes of values (in degrees) of ϕ , ψ , ω , χ dihedral angles of all tachykinin peptides lowest energy

conformations under MD simulations in vacuum and water are compared. The deviations of ψ for Arg1 by $\pm 20^\circ$ from its optimal values are allowed in all calculated structures in vacuum and water environment. The low energy changes of χ_1 for Arg1 from 182 to 88° are possible. It is shown that the Phe7 and Phe8 side chains are close to the minima of the torsional potential. The deviations by $\pm 20^\circ$ from minimal values are possible for χ_1 angle. The rotation of the χ_2 angle for Phe7 and Phe8 is considerably limited due to the effective interactions between the Phe7 and Phe8 amino acids. The mobility of the backbone and side chain of the Leu3 is more restricted as compared to preceding residues of molecules in vacuum as well as in water. In contrast to water simulations, where the ϕ angle for Lys3 may be changed by retained, generally the bond stretching frequency of water, the trajectory has to be recorded at an interval no larger than 4 femtoseconds. MD simulations show that the molecule backbone can adopt only a limited number conformations while the side chains of the residues may populate different rotamers. A large flexibility of the Arg1-Pro4 amino acids sequence was observed in vacuum in contrast to water simulation. The Gln5-Met11 fragment was found to be rigid in the conditions studies. Changes in intramolecular energy during simulations in water were negligible; they did not exceed 10-15 kJ/mol for molecule. At the same time, the molecule interaction energy was much higher due to the flexibility of the Lys3-Gln6 part of the peptides. Interactions between aromatic side chains of the Phe7 and Phe8 amino acids make the largest contributions to the global energy of the simulated molecule. Undoubtedly this contribution is overestimated in the vacuum approximation. MD simulations for physalaemin and uperolein during 300ps indicate that ψ angle for pGlu1 have a noticeable conformational flexibility. All side chains angles of Lys 6 were seen to be well-defined around 180° throughout the runs. The run with the all low energy starting structures had an initial angles for Asn5 – Phe7 change its only around -10° and 10° . The conformational changes during the MD simulations in vacuum and in water are shown on Fig.1. As a result sinking the global structure of three molecules in the box with water molecules was received the final optimized structure of peptide at the during the MD simulation (Fig.1). The mobility of the Asp3- Lys 6 amino acids stretch is considerably limited. So, the flexibility of residues in the 5th and 7th positions is limited by 10° as compared to the preceding part of molecule. This fact can be explained due to the important role of these residues in the formation of β -turn. Each angle varied about a single value, close to one of the set of possible angles calculated from molecular mechanics energy minimization [12].



Fig.1. Preferred spatial structures of the physalaemin ,uperolein and Substancer P molecules.

IV. CONCLUSION

We have carried out detailed analysis of the flexibility of the tachykinin molecules by employing the molecular dynamics method. The foregoing results and discussion lead to the following conclusions: (I). molecular dynamics simulations in vacuum as well as in aqueous solution confirm the considerable flexibility of the 1-4 sequence of substance P; (II). the α -helical conformation on 5-11 segment of peptide was more stabilized in vacuum, with the predominant hydrogen bonds than the extended conformations; (III) the similar molecular dynamics simulations for physalaemin and uperolein indicated that relatively high stability of the low-energy conformations resulted not only from nonvalent interactions between residues but also from hydrogen bonds networks; (IV) the β -turn conformation at the 3-6 segment were more stabilized in vacuum and provide optimal nonvalent interactions between residues. The determined structures of these tachykinins may be used as the basis for the design of further peptidic selective antagonists.

This work have been fulfilled in the frame of collaboration treaty of the Qafqaz and Baku State Universities.

REFERENCES

- [1] Euler V.S., Gaddum J.H., *J. Physiol.* (1931) 72, 74-87.
- [2] Chang M.M., Leeman S.E., Niall H.D., *Nature New Biol.* (1971) 232, 86-87.
- [3] Guard S., Watson S.P., *Neurochem. Int.* (1991) 18, 149-165.
- [4] Chassaing G., Convert O., Lavielle S. *Eur. J. Biochem.* (1986), 154, 77-85.
- [5] Convert O., Duplax H., Lavielle S., Chassaing G. *Neuropeptides* (1991), 19, 259-270.
- [6] A. Anastasi, V. Erspamer, *Br. J. Pharmacol. Chemother.*, 1962, v.19(2), p.326-336.
- [7] V. Erspamer, A. Anastasi, *Experientia*, 1962, v.15, N18, p.58-59.
- [8] IUPAC-IUB, *Biochem. J.* (1971) 121, 577.
- [9] T. Wymore, T.C. Wong, *Biophysical Journal*, 76, (1999), 1199.
- [10] C.A. Maggi, *General Pharmacology*, v.26, iss.5, (1995), p.911-944.
- [11] I.S. Maksumov, L.I. Ismailova, N.M. Godjaev, *J. Struct. Khim.*, vol. 24, 1983, pp.147-148.
- [12] G.A. Agaeva, N.N. Kerimli, N.M. Godjaev, *Biofizika*, vol. 50, 2005, pp. 203-214.
- [13] W.F. Van Gunsteren, P.K. Weiner and A.K. Wilkinson in *Computer Simulation in Biological Systems* (ESCOM Science), 1993.
- [14] J.A. McCammon and S.C. Harvey in: *Dynamics of Proteins and Nucleic Acids* (Cambridge Univ. Press, New York), 1987.
- [15] S.N. Rao and P.A. Kollman. *Prog. Natl. Acad. Sci. USA*, 1987, 17, 6883.
- [16] M. Billeter, A.E. Howard, I.D. Kuntz and P.A. Kollman. *J. Am. Chem. Soc.*, 1998, 110 8385.
- [17] S.W. Chiu, J.A. Novotny and E. Jakobsson. *Biophysical J.* 1989, 64, №1, 98.
- [18] W.D. Cornell, P. Cieplak, C.I. Bayly, I.R. Gould, K.M. Merz, et al. *J. Am. Chem. Soc.*, 1995, 117, 5179.
- [19] N.L. Allinger, V. Yuh. *QCPE-395*, Indiana Univ., 2000.

Application of Computer Technologies in Investigation of Spatial Structure of Peptide Molecules

G.A.Akverdieva

Institute for Physical Problems, Baku State University, Z.Khalilov Str.23, AZ-1148,

Baku, Azerbaijan, E-mail: HagverdiguLnara@gmail.com

Abstract- In this work the possibilities of the theoretical study of the spatial molecular structure using a computer technology advances are demonstrated on the example of the atypical opioid peptides - hemorphins. The calculations of the conformational profiles of these molecules were carried out by the theoretical conformational analysis method, the electronic structure of the conformations was investigated by quantum-chemical method, the molecular dynamics of peptide molecules was investigated using the demo version of HyperChem program. On the base received results and data of biological testing the computer modeling and the comparison of the putative bioactive conformations of the investigated peptide molecules were conducted. The characteristics of the active center of hemorphins were assessed independently.

Key words: spatial structure, conformation, hemorphin peptides, theoretical conformational analysis, molecular dynamics, electronic structure

I. INTRODUCTION

The development of the ideas about the mechanism of action of the peptide molecules, understanding of the causes of the activity and selectivity of the drugs that interact with receptors are possible due to the structural and functional studies on the molecular level that can not be achieved solely on the basis of experimental methods. Thus, a knowledge about the spatial structure of the crystal obtained by X-ray analysis or in the solution using physical-chemical methods does not provide to complete the information about the biologically active conformations of the peptides. The physiological activity of peptide molecules is directly related to their particular spatial structure with its characteristic electronic parameters and dynamical conformational properties that play an important role in the enzyme-substrate interactions. Currently, using the different theoretical calculation methods, the recent advances in

computer technology, including programmes with a graphical representation of the spatial structures allow researchers to construct the various models of the peptide molecules. Thus, it is possible to obtain the spectrum of low-energy conformations of the molecule corresponding to its biologically active states by the method of the theoretical conformational analysis, to investigate the electronic structure of each conformational state of the molecules by quantum-chemical methods, to reveal the equilibrium properties and the kinetic behavior of the molecules in the vicinity of the specified minimum by the molecular dynamics method. In this work the possibilities of a theoretical study of the spatial molecular structure with the using of the computer technology advances are demonstrated on the example of the hemorphins molecules.

Hemorphins are atypical endogenous opioid peptides [1]. In organism these peptides are generated by consecutive enzymatic hydrolysis of the β -chain of the blood protein hemoglobin. Hemorphins have been identified as naturally occurring peptides in brain, plasma and cerebrospinal fluid. These peptides participate in regulation of the activity of some biomolecules and demonstrate a number of the diverse physiological activities. The knowledge of spatial structure and conformational properties of hemorphins is essential for elucidation of molecular mechanism and specify of their action.

II. METHODS

The calculations of the conformational profiles were carried out using the developed in the Institute for Physical Problems of Baku State University universal computer program for the calculation of biomolecules, written in FORTRAN [2]. The investigations were carried out using the theoretical conformational analysis as described in [3]. The conformational potential energy of the peptide molecules is

calculated as the sum of the independent contributions of nonvalent, electrostatic, torsional interactions and hydrogen bonds [4, 5]. To find the minimum energy conformation the method of conjugate gradients was used. The electronic structure of the conformations was investigated by AM1 quantum-chemical method and molecular dynamics of peptide molecules was investigated using the demo version of HyperChem program (<http://www.hyper.com>) [6, 7].

III. CALCULATIONS AND RESULTS

A. Hemorphin peptides

The energy and geometrical parameters of stable states of hemorphin-4, VV-hemorphin-4, spinorphin, LVV-hemorphin-7 opioid peptid molecules, including from 4 to 10 aminoacid residues (Leu-Val-Val-Tyr-Pro-Trp-Thr-Gln-Arg-Phe) are calculated [8-11]. The investigated peptides are active molecules, and the sequences of the small peptides are the parts of larger peptides from this family. It is established that the investigated sequences have determine conformational mobility in space; the spatial structure of their sequence can be described by several types of the backbone, though their energy is very sensitive both to shape of backbone and to positions of the side chains of the amino acid residues. It was established that the conformational state of the central fragment Tyr-Pro-Trp-Thr plays the important role in the energy differentiation of the structures and in the formation of the stable states of the hemorphin peptides. It is significant that the the turn of the polypeptide chain was revealed on the segment Pro-Trp of the physiologically active tetrapeptide part Tyr-Pro-Trp-Thr in the optimal structures of all investigated hemorphin peptides. In such conformations Tyr and Trp are in *cis*-positions in respect to Pro, due to that the planes of the aromatic rings of their side chains form 90^0 or 0^0 , that provides the maximal approachment and interactions of atom groups. The conformational and molecular dynamics of LVV-hemorphin-7 is also investigated The mobility of the functional groups of the molecule is estimated and the conformationally rigid and labile segments of this peptide are revealed [12,13]. These studies also showed that the central part Tyr4-Thr7 is conformationally rigid in comparison with the terminal segments of of this molecule.

The conformational profiles of biologically tested analogues [Ala4]- and [Ala6]- LVV-hemorphin-7 are investigated and the role of the substituted residues in the formation of the spatial structure of this peptide molecule was revealed [13]. The calculations showed that at replacement of tyrosine by alanine the turn on the fragment Tyr-Pro-Trp-Thr remains unchanged. We can conclude that as a result of this substitution the loss of activity is due to the loss of chemical groups, which are localized in the side chains of tyrosine, but not due to conformational factors. The replacement of tryptophan by alanine leads to a deformation of the turn at a mentioned tetrapeptide resulting in a chain of tripeptide part of

Tyr4-Trp6 takes an elongated shape. One can conclude that the side chain of tryptophan helps to stabilize the turn on the segment Tyr4-Thr7 of this peptide. Thus, the aromatic amino acid residues Tyr and Trp are involved in the physiological activity of this molecule. On the basis received results a computer modeling of the spatial structures of bioactive conformations of the investigated hemorphins molecules was conducted [14], the results of collation of these structures are illustrated In Fig.1 and Fig.2. As can be seen from the presented figures the conformations of the central tetrapeptide part Tyr-Pro-Trp-Thr of hemorphin molecules are similar. Using the algorithm of the estimation of the geometrical resemblance of the pair of conformations it is realized that root-mean-square deviations of the coordinats of backbone atoms and of the distances between the atoms of this fraqment in investigated hemorphin peptides are small. These results show that the tetrapeptide part Tyr-Pro-Trp-Thr of hemorphin molecules is conservative, resistant to conformational changes.

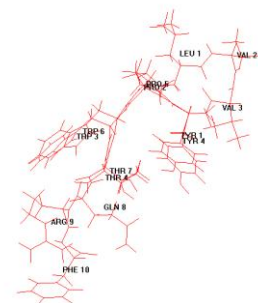


Fig.1. The comparison of the global conformations of LVV-hemorphin-7 and hemorphin-4 peptides as a result of superimposition of the coordinates of nonhydrogen atoms

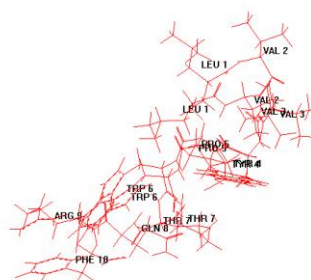


Fig.2. The comparison of the global conformations of LVV-hemorphin-7 and spinorphin peptides as a result of superimposition of the coordinates of nonhydrogen atoms

B. Active center of hemorphins

The structure-functional investigations of hemorphins allow to consider Tyr-Pro-Trp-Thr tetrapeptide fragment as active center, playing an important role in the formation of the bioactive conformations of these peptides and providing the specificity of their action. For more detailed study the dynamical properties and electronic structure of the active center of hemorphins have been studied separately [15,16]. It is found, that the temperature changes are reflected on the dynamics of atoms of the side chains of tetrapeptide Tyr-Pro-Trp-Thr. Unlike Tyr aminoacid residue Trp is more dynamic in vacuum, than in water environment. It can be concluded that a certain number of water molecules connect with the atoms of side chain of tryptophan and limit the movement of this residue as well as of the whole peptide molecule. It can be assumed that the side chain of this residue may be involved in enzyme-substrat interactions. On the basis of values of the effective charges on the atoms, of the dipole moments, of the distribution of the electronic density and of the electrostatic potential the electron-conformational properties of this tetrapeptide were studied. At optimization of the electronic energy as zero approach the configurations of nuclears, corresponding to geometries of the optimal structures of the tetrapeptide fragment Tyr-Pro-Trp-Thr were considered. The calculation model of active center with indication of atoms is shown on Fig.3. We estimated such quantitative parameters characterizing the electronic structure of this fragment as the electron population and the orbital energy. These parameters proved to differ for the considered conformations of the tetrapeptide. It was revealed that each of the investigated conformations of this tetrapeptide has the specific distribution of electron density, which reflects on the values of effective atomic charges of the functional residues. The analysis of values of charges has allowed to conclude that the conformational transitions in low-energy intervals are accompanied by fluctuations of density of the charges, not exceeding 20 percent. There are differences in the values of the charges on certain groups of atoms, both main and side chains of residues, that dictated by specificity of their relative positions in each structure. So, changes are mainly in the charges of the atoms of the side chains with aromatic rings - Tyr, Trp, and also of the atoms of the main chain of Thr, spatially approached with them.

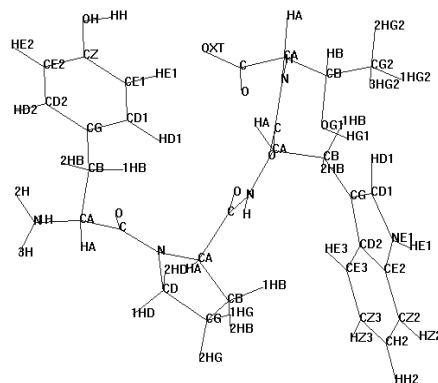


Fig.3. Calculation model of active center of hemorphins

The calculations have shown that the global conformation of active center, having the turn on the segment Pro-Trp is characterized by optimal electronic structure. This structure has much smaller dipole moment than other conformations due to the uniformity of the distribution of the electronic density, of the negative and positive charges. This structure is characterized also by smaller value of energy of the higher filled molecular orbitals which correlates with potential of ionization of a molecule and characterises its electron-donor properties. Thus, it is possible to believe that global conformation has considerably less expressed chemical reactionary ability, so also weak electrodonor properties, than other structures that is probably connected with high degree of activity. The analysis of a picture of distribution of electrostatic potential in this structure of the tetrapeptide has revealed a strongly pronounced zone with the raised electronic density on which the attack of electrofil reagent can be directed (Fig.4). We will notice that this structure of the tetrapeptide forms a loop in the optimal conformations of the investigated hemorphin peptides.

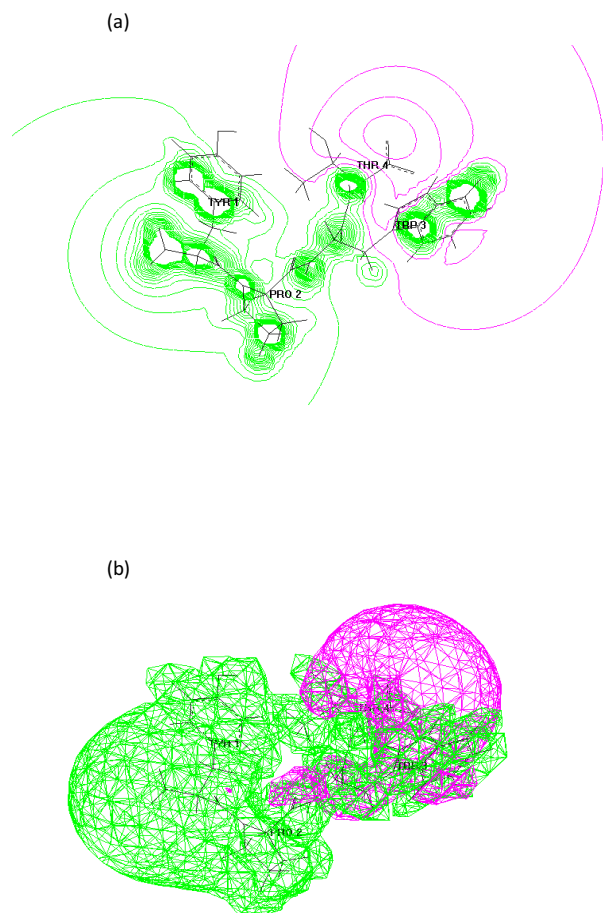


Fig.4. The 2-D (a) and 3-D (b) distribution of the electrostatic potential in the global conformation of active center of hemorphins

IV. CONCLUSIONS

Thus, the results of study hemorphins by molecular mechanics, molecular dynamics, quantum-chemical methods confirmed the same mechanism of differentiation of conformations. The received results allow to assess the geometrical characteristics of bioactive conformation of active center of hemorphin peptides (see Table). The investigations of electronic and conformational properties of active center of hemorphins can appear an important basis for additional correlation of the spatial structure and structure-functional interactions of peptides from hemorphins family and may be useful in the design of a new synthetic peptides with selective activities.

TABLE
GEOMETRY PARAMETERS of BIOACTIVE CONFORMATION of
ACTIVE CENTER of HEMORPHINS
(in degrees)

Tyr	φ	-115
	χ_1	58
	χ_2	89
	χ_3	180
	ψ	149
Pro	ω	176
	ψ	-61
Trp	ω	180
	φ	-159
	χ_1	179
	χ_2	83
	ψ	144
Thr	ω	180
	φ	-135
	χ_1	-62
	χ_2	180
	ψ	113

This work has been fulfilled in the frame of collaboration treaty of the Qafqaz and Baku State Universities

REFERENCES

- [1] Zhao Q, Garreau I, Sannier F, Piot JM. Opioid peptides derived from hemoglobin: hemorphins. *Biopolymers*. 1997, Vol.43, Issue 2, p.75-98
- [2] Максумов И.С., Исмаилова Л.И., Годжаев Н.М. Программа полуэмпирического расчета конформаций молекулярных комплексов на ЭВМ // *Ж. струк. химии*, 1983, т.24, с.147-148
- [3] Godjajev N.M., Akuz S., Akverdieva G., *J.Mol.Struct.* 1997, v.403, p. 95-110
- [4] Попов Е.М. // *Int.J.Quantum Chem.* 1979. v. 16. p.707-737.
- [5] Попов Е.М. *Структурная организация белков*. М. Наука. 1989. 352 с.
- [6] Шайтан К.В., Сарайкин С.С. *Метод молекулярной динамики*. [http:// www.moldyn.ru](http://www.moldyn.ru)
- [7] Allinger N.L., Yuh Y. QCPE 395, *Quantum chemistry program exchange*, Indiana Univ., Indiana, 1982
- [8] Qocayev N.M., Haqverdiyeva G.Ə., Nəbiyev Ə.M. Hemorfinlərin bioloji aktiv hissəsinin konformasiya tədqiqi. *Journal of Qafqaz University*, 17 Fall, 2006, p.107-111
- [9] Nəbiyev Ə.M., Haqverdiyeva G.Ə., Qocayev N.M., VV-hemorfin-4 peptidinin fəza quruluşu. *Journal of Qafqaz University*, 17 Fall, 2 006, p.63-69
- [10] Nəbiyev Ə.M., Haqverdiyeva G.Ə., Spinorfinin fəza quruluşunun tədqiqi. *Journal of Qafqaz University*, N.19, 2007, p.87-94
- [11] Nəbiyev Ə.M., Haqverdiyeva G.Ə, Qocayev N.M. LVV-hemorfin-7 peptidinin fəza quruluşunun tədqiqi. *Journal of Qafqaz University*, N.23, 2008, p.72-78
- [12] Nəbiyev Ə.M., Haqverdiyeva G.Ə. LVV-hemorfin-7 peptidinin konformasiya dinamikası. *Bakı Universitetinin xəbərləri, fizika-riyaziyyat elmləri seriyası*, № 1, 2010, s.151-161
- [13] Nəbiyev Ə.M., Haqverdiyeva G.Ə., Qocayev N.M. LVV-hemorfin-7 peptidinin quruluş-funksiya xassələri. *International. AMEA-nın Fizika jurnalı*, № 2, cild XVI, seriya: AZ, 2010, c.70
- [14] Axverdiyeva G.A., Nəbiyev A.M., Godjajev N.M. Теоретический подход к моделированию биологически активных конформаций геморфинов. *6-я Всероссийская конференция "Молекулярное моделирование"*, 8 - 10 апреля, 2009 г., стр.46
- [15] Nəbiyev A.M., Axverdiyeva G.A., Godjajev N.M. Исследование электронной структуры пептида геморфин-4. *Journal of Qafqaz University*, N 30, Volume 2, 2010, p.25-32
- [16] Akverdieva G.A., Nəbiyev A.M., Godjajev N.M. Computer modeling of active center of hemorphins. *V International Conference on Application of Information and Communication technologies (AICT2011)*. 12-14 October 2011, Baku, Azerbaijan, p. 689-691

Imitation modeling of competitive market equilibrium

Khatuna Bardavelidze

Department of Computer Engineering
Georgian Technical University, GTU
Tbilisi, Georgia
bardaveli_x@yahoo.com

Avtandil Bardavelidze

Department of Computer Technologies
Akaki Tsereteli State University, ATSU
Kutaisi, Georgia
bardaveli@yandex.ru

Abstract — This paper is about questions of the computer modeling of the influence hypothesis for demand and supply at the dynamic of market equilibrium prices. By result of competitive market is worked out mathematical model, on which base composed functional scheme. By means of functional scheme of market equilibrium and mathematical model is composed system computer model Simulink as block-scheme. Automatization of experiments is carried out by MatLab program. With changing of model parameters is conducted analysis of influence of demand and supply at the market equilibrium.

Keywords- *competitive market; supply and demand; imitation model; mathematical model; computer and functional model; market equilibrium;*

I. INTRODUCTION

Imitation modeling of competitive market presents one of the most important factor for research economical processes, theoretical and practical interests of building and researching their mathematical models, which describe market price dynamics with great adequacy, volume of delivery and sales goods at the market in depending on relation demand and supply, competition of goods and sellers, discipline of goods delivery, marketing policy and strategy of market participant and different factors which have effect on the stability of state market equilibrium and character of market transient processes.

For solution standard economical tasks, such as searching equilibrium of demand and supply is reasonable to use imitation modeling. It allows us to simulate system behavior in time. Though advantage of this is that by time in model can be controlled: to delay in case with fast processes and accelerate for modeling systems by slow variability.

Every system tends to achieve and keep its equilibrium state which is character for microeconomic systems. Although, as its functioning is provided by people activities, imposed obligation on them and to the multidirectional interests, market balance does not achieve spontaneity, because it has specific laws and conditions.

The founder of prices theory Alfred Marshal considered that most of economic process can be explained by terms of market equilibrium prices. The price ascertaining on the

product happens during interaction of demand and supply. Usually, the lines of demand and supply is drawn manually in relation to goods price, they move lines, change their characteristics and observe new equilibrium points. All above mentioned can be realized by using of MatLab program package, by forming and modeling of imitation model for market equilibrium state [1].

Thereby, market equilibrium is a position of state, in which volume of demand is equal to volume of supply. Demand and supply get balanced under influence of competitive market, in consequence that price accords to quantity goods, which company is agree to sell, and customers are willing to pay money. In the equilibrium market price coincide economical interests as sellers as customers.

II. PROBLEM DEFINITION

Transition to the market economical interrelation formed different kind of problems, from one of which presents mechanism investigation of achievement market balance state of products demand and supply at the competitive market. To solve this problem is necessary to implement computing and as result to receive the graph with picture of crossing the lines of demand and supply at the balance point [1, 2].

Balance achievement on the competitive market is so important for stable and effective development of company that this problem cannot stay without attention, because market equilibrium achievement means proportion between product and supply, offer and demand, in manufacturing expenses and results.

Goal of this paper presents to form and analysis imitating model at the competitive market for equilibrium of goods demand and supply, to compose functional scheme and its computer model. For realizing this purpose is used MatLab program package with the meaning of visual modeling tools – Simulink, which graphically represents market equilibrium state of demand and supply.

Generally, at the company level determined problem we can imagine as following: let us assume that owner intends to invest fund for creating company, which will produce goods and make its realization at the market. He interested what influence will have goods price during the change of produce size. It is necessary to receive the answer, in what condition the price of goods will be stable, taking into consideration while manufacture increasing at the market demand of the

production decreases and discounts its price. This kind of task can be solved by way of imitation modeling.

III. MATHEMATICAL MODEL

In the literature [1, 2] described several modification models, which all have definite identical properties. In them usually supposed that *demand* at certain product in the given space of time depends on price and other factors on this section. As concerning *supply*, it is determined by prices previous period of time. In addition that supposed that market is always in conditions of local equilibrium. Historically that model is called as „cobweb model“.

There exist four variants of this model: deterministic, probabilistic, model with teaching and model with supplies. In deterministic model absented calculating influence of random factors. In probabilistic model includes influence on the demand unforeseen variation of preferences and consumer income, also other random factors which have influence at the quantity demand.

Supply at the previous period of time also supposes subjected of influence random factors. They reflect influence of variation technology and efficiency of industrial process. At last, condition of local equilibrium means confluence of demand and supply with accuracy to some random variable.

In model with teaching supposed that suppliers take into account formed tendency of changing prices and by accounting this they plan production output at the next interval of time.

In last two models prices are established at that level in order to provide local equilibrium of market just at the expense of current manufacture and none reserve of product doesn't create because for example, products spoil quickly. In model with supplies is included additional participants group of market mechanism, which we can call as „businessmen“. They hold stocks and organize commerce.

For our case is more convenient probabilistic model with teaching. It is supposed that demand at the time section of $T-m$ linearly depends on current price and except that demand is subjected to random scatter. Thereby, dependence function of demand at the price is following [2]:

$$D_{md} = D_0 - K_d * P_{rc} + U, \quad (1)$$

where D_{md} – is demand in current interval T time; D_0 – demand with zero price; K_d – curvature of demand line; P_{rc} – enable to be determined price at the interval of $T-m$; U – is random variable with given distribution law. Supposed, that demand symmetrically oscillates to relative average value, which is determined by constant coefficients of linear equation. So we can select normal distribution with zero mathematical expectation and with given average quadratic deviation.

Supply at the current interval of time linearly depends on the price, not current but by self representative some price combination on two previous intervals of time. In common case it can be average price. Therefore, for calculating of supply uses follow dependence:

$$S_{pl} = S_0 + K_s * P_{rc} + V, \quad (2)$$

where S_{pl} – is goods supply at the interval of $T-m$; S_0 – supply during zero price; K_s – curvature of supply line. P_{rc} – enable to

be determined price at the interval of $T-m$ time; V – is random variable with given distribution law.

At the end condition local equilibrium of market can be written, as

$$S_{pl} = D_{md} + W \quad (3),$$

where W – is random variable with given distribution law. Random variable - W is characterized by zero mathematical expectation and by average quadratic deviation. By consideration that model is simplified, step of definition problem and step building of conceptual model coincided. By substituting expression D_{md} and S_{pl} in (3) and solving equation relative to P_{rc} , we get:

$$S_{pl} = D_0 - K_d * P_{rc} + U + W, \quad (4)$$

$$P_{rc} = (D_0 - S_{pl}) / K_d - U - W. \quad (5)$$

Task of modeling included in researching of influence system's parameters on the character price dependence from time.

IV. COMPUTER MODELING

On the basis above presented market equilibrium model analysis, is worked out functional scheme, fig. 1.

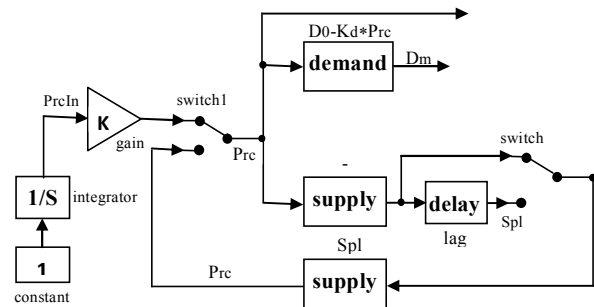


Figure 1. Functional scheme of market equilibrium

On the fig.1, demand is presented by standard block $DmdFn$, which calculates at the scheme entrance given demand signification in relation to price. Supply is presented with three standard blocks. Directly, the function of amount selling goods of dependence on price is realized by $SplFn$ block, which calculates the significance value of given supply on the entrance in relation to price. Lag - block means delay of goods supply at the market. Seller presents Spl amount of goods which is determined basis on past time interval of price. By $SplFn1$ is imitated receiving decision of supplier with the current demand price. He consents to sell all goods by the price which dictates the demand line. With mentioned block realizes inverse function of demand and calculates price of Prc , by which supplier can sell all Spl goods. The parameters of block - $SplFn1$ and demand - $DmdFn$ are the same.

On the fig.1, the blocks of constant, integrator and gain give us the prices value for building statistical characters of demand and supply. The meaning of switchers is to conduct modeling in different regime: 1) to build the function of demand and supply in relation to price and 2) to build the graph of market transient process in equilibrium state [2].

From presented market equilibrium functional scheme on the fig.1 and from mathematical model, we can compose computer model of system as block–scheme – SIMULINK [3]. It consists standard functional blocks of control system and objects, fig.2. Transient computer models of mathematical functions are presented by blocks. The marks in blocks perform the analysis expression formulas of transfer function.

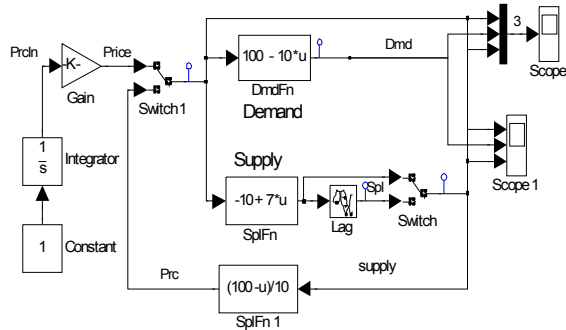


Figure. 2 Block scheme of market equilibrium imitation model

At the fig. 2, the blocks parameters are following: $U=P_{rc}$, $D_0=100$, $K_d=10$; $S_0=10$, $K_s=7$. Automatization of experiment is executed by Matlab program with m-file commands of experiment's control:

```
%Market equilibrim price simulation
open_system('DmdSplEqu')
sim('DmdSplEqu')%Write Vars into WS from Scope
%Plot Static features
plot(ScopeData(:,2),ScopeData(:,3:4))
hold on
grid
pause(5)
%2.Simulate price dynamic
sim('DmdSplEqu')
%3. plot price Web graphics
for i = 2:11
    line([ScopeData(i-1,2) ScopeData(i,2)],[ScopeData(i,4)
    ScopeData(i,4)])
    line([ScopeData(i,2) ScopeData(i,2)],[ScopeData(i,4)
    ScopeData(i+1,4)])
end
hold off
```

The windows of blocks Scope and Scope1, which presented on the fig. 2 are demonstrated on the fig. 3 and fig. 4. They reflect the graphs of price changing of demand and supply in relation to time, but on the fig. 5 – the graph of equilibrium transient process of competitive market, its called as „cobweb graph“, which will be built by above showed experiment's control Matlab program with m-file commands.

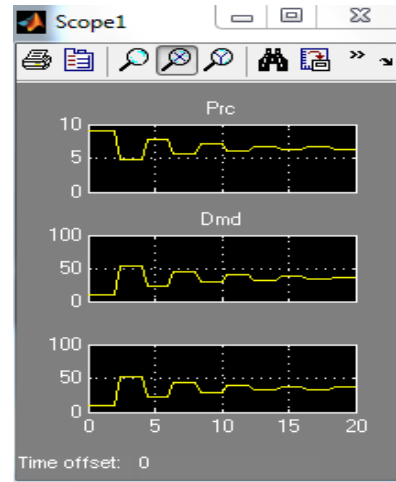


Figure 3. Prices changing in time of demand and supply

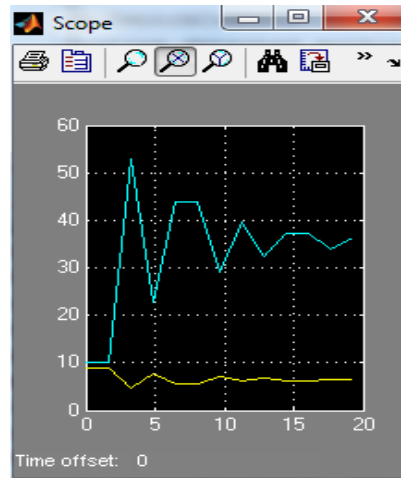
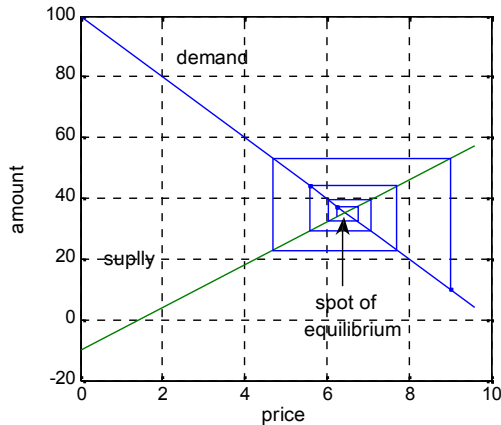


Figure 4. Price changing of demand and supply in scope window

“Cobweb” graph for market equilibrium model of movement is presented on the fig. 5. For learning the transient process of market equilibrium in Simulink window we set switcher in bottom position and in Scope 1 window we observe graphs of changing (Prc, Dmd) characteristics in time, fig. 3 and graph of price movement at the fig. 5. Scope graph is presented at the fig. 4.

In order to analysis the influence of demand and supply at the market equilibrium, on one side, we change value of D_0 parameter in blocks DmdFn, SplFn1 and observe price movement to new equilibrium, on the other side we change S_0 parameter value in blocks SplFn and observe movement to new equilibrium. But for analyzing influence of curvature line at the market equilibrium of demand and supply, on one side we change value of K_d parameter in DmdFn and SplFn1 blocks and observe movement of prices to new equilibrium, on the other side, we change K_s parameter value in SplFn block and connect to the new movement of equilibrium.

Figure 5. Equilibrium "Cobweb" graph of prices movement at the market
web movements for market equilibrium



V. CONCLUSION

Basis on above conducted investigation and analysis is established that during increasing of K_s parameter the oscillating indicator raises in system, but when $K_s=K_d$, the oscillating indicator don't change in time, steady equilibrium doesn't achieve and the system locates in steady oscillating regime. When $K_s>K_d$, system isn't stable and the process is spreading. That is oppose of real economics and approve that model is rough.

On basis of analyzing established that in changing of model parameters the oscillating indicator raises in system, steady equilibrium doesn't achieve. In often occurrence system isn't steady and model is rough. To compose imitation model of market equilibrium and use the receiving value by result of modeling importantly simplifies the solution of searching task for demand and supply at the market equilibrium.

Computer model of market equilibrium was built by result of conducted imitation modeling. Received point by crossing of curvature demand and supply presents equilibrium price while whole output productions will be realized; Graphics results of modeling completely conform to necessary criteria of market equilibrium, goods demand and supply at market and prove their adequacy.

REFERENCES

- [1] A. I. Dobrinin, L.C. Tarasevich, "Economics Theory," 3rd ed., Sankt-Peterburg: Piter, 2001, p.544 (in Russian)
- [2] N.Snetkov, "Simulation modeling of economical process", tutorial, Moscow: Center of European Public Institute, 2008, p.228 (in Russian)
- [3] O. Beucher and M.Weeks. Introduction to MATLAB & Simulink: A Project Approach, 3rd ed., Hingham, MA: Infinity Science Press LLC, 2008, p.404.

Methods of evaluating students In different countries

Mehdi H.Soltani, Alovzat Q.Aliyev
Institute of Information Technology of ANAS, Baku, Azerbaijan
soltani466@yahoo.com; alovzat@iit.ab.az

Abstract

If there isn't educational evaluation in educational system, education will lose its position. So in this essay is tried to examine different educational ways of students in different countries, there different ways are improvement, recognition, portfolio, project, final, combine, monitoring, accumulative evaluation. In conclusion by using these ways, we can show results of educational system and get desired results and examining reasons of growth and progress and lack of growth and progress, if these evaluations conduct correctly.

Keyword

Project; Continuous; Communicative; Diagnostic; student
Condensation; Observational; Condensation

Introduction

Introduction about the word evaluation :

The word evaluation in its broad meaning is about the process according to which and based on some criteria, judgment about a person is made. Evaluation is a process which is done at the end of each period or is a systematic process for doing a research and interpreting of the information in order to see whether the required purposes have been fulfilled or not?

Evaluation is one of the important and indispensable principles of education. We should bare it into our minds that education is consistent and firm process, all parts of which are organized in a way to fulfill a specific goal. At the end of each educational period the interested teachers assess their students progress and change to see how much they have approached the required purposes. This process is called evaluation which includes important and long term purposes in modern views, unlike the past. The ultimate goal of evaluation is improving students learning. When the evaluation is not appropriate and right it may undergo some deficiencies such as lack of students interest, material and humanistic investment wasting, increasing the number of failing students, hindrances in the students emotional progress and fading of students creativity and so on. So students observation in the class, laboratory and other educational situations, using informal conversations and interviews, analyzing academic activities of students, conducting various tests, and so on are considered as important teachers there are different kinds of evaluation, in this article 8 methods of evaluation are described and compared which are as following: Continues evaluation-summative -portfolio evaluation- project evaluation- final evaluation- triangulation evaluation - observational evaluation- condensation evaluation.

In many countries at least one of these methods is applied

1- Continuous evaluation:

This evaluation is applied in the middle of teachers educational activities and the students learning are being constructed. Teachers can use this methods in order to solve students problems, misunderstandings and learning deficiencies while the educational activities and attempts are in the process. The important characteristic of this is its seeking and vast scope. And teachers are expected to take further steps out of traditional assessment in order to give a general image of students characters. The continues evaluation starts form the very beginning of the educational process and help teachers to recognize the students abilities and be aware of their weak and strong learning points. Continues feedback to students is one of the important factors of this method. One of the important goals of continues evaluation is for teachers to recognize students learning weak and strong points, find appropriate ways and arrange their forthcoming teaching considering those factors. Teachers themselves are in the process of experience and by this method they can recognize students interests and assess the activity and content level of learning. To put it in nutshell, continues evaluation is judgment about the gathering of information through educational activity which is done in order to decision making and improving learning process. In United state of America and many European countries such as Finland, Germany and Italy this method is used.[1]

2- Diagnostic evaluation:

This kind of evaluation is done usually through the process of learning in order to recognize students learning deficiencies in an academic course. In addition to recognize students problems, this method should also present to the teachers appropriate ways and procedures to solve these problems. Diagnostic evaluation helps teachers to match course content and educational method with the students previous learning and their needs. It also helps teachers to get familiar with appropriate learning experiences. Teachers based on the result of this method fill the gap between students ability and new subjects requirements. (In Malaysia this method is used) .[2]

3-Portfolio evaluation(Folder):

Portfolio evaluation is conducted regarding students overall activities in or outside of the class. This method requires archive of document and certificate related to students progress through the field shows their real improvement. In this evaluation students learn how to evaluate themselves through the thinking about their sample works and see their progress themselves. Teachers give them necessary feedback by designing appropriate questions out of students activities

Portfolio let students know and correct their mistakes. This enables students not to be afraid of their mistakes and help them to risk and correct their mistakes which in turn help students to share their involvement in the process. (This method is used in Australia, England, Scotland, Ireland)[3]

4- Project evaluation:

This evaluation consists of activities which are based on course purposes and a bit complicated than common activities. In contrast to education system plans and activities which have nothing to do with student design and don't allow them to be creative. In this evaluation students are free to make decision about their academic affairs and they improve this ability in them selves. In this method students are supposed to gather information for their project and for so doing they do many activities such as choose a topic , problem recognition, research design, Work sharing, perseverance using different teaching use of problem solving, continues work assessment while doing it, providing record using scientific procedures. In this method students also interview many people with different occupations in order to gather information and get familiar with their opinions. They also refer to many valid references. The role of teachers is as facilitators and guides and students have the opportunity to gain skills of better life affairs and facing their problems. The project is based on course syllabus and aims at discovering and recognizing students educational failure in particular courses. This project helps teachers to know the students problems and try to solve them. (This method is used in eastern countries in Asia like Japan and China)[4] .

5- Communicative or final evaluation:

This evaluation which is also called summative assessment, aims to measure or summarize what students have grasped and typically at the end of a course or unit of instruction. The final goal of this method is to give score to students and see how successful teachers were with regard their teaching process. According to Pasha sharifi: Summative evaluation occurs at the end of course or unit of instruction and aims at specifying levels of knowledge, skills, ability, enhanced point of view in the process of course of instruction based on specific plans and educational system purposes. For example it's about whether students have learned whatever they were supposed to or not? Summative evaluation usually occurs at the end of course of instruction and teachers usually use this method to assess what students have learned through this period at it aims at giving score and decision making about the students qualifications to go to the higher levels and so on. The way these evaluation is distinguished with formative evaluation is how to apply the final result. The final results of this method has nothing to do with next levels and the responsibility to discern students problems through the process of learning and solving them is on the shoulder of formative assessment. The aim of summative assessment is to specify the amount of students learning at the end of course of instruction and to judge about

the overall educational system plans and design which is reported to the school office and parents in the form of qualitative or quantities score. Although in Canada educational system is in the form of state , but in many state of this country, summative evaluation is used.

6. Synthetic evaluation

This evaluation is consist of two qualitative and quantities method and is a comprehensive one. Using result of two methods, this evaluation has a better and enhanced assessment and this process which is called triangulation is one of the core advantages of combined evaluation. The combination of two qualitative and quantitative method can be effective in constructing all the evaluation measures in all of the methods to gather the data.[5]

7. Observational evaluation:

Using list of observations in educational system is one of the modern and novel method by which teachers can comment on students performance and learning in different fields of knowledge, skill and view point. The list of observation is so flexible and is designed according to each teacher expectation from student which may be different form other teachers list. The list of observation can alarm teachers in many fields of students need to get essential steps for paving the way for them before its too late. By using this method teachers can help students to enhance their conscious behavior such as perseverance, listening, thinking flexibility, super recognition, caring action and so on.

8. Condensation evaluation:

This kind of evaluation is so common in our country schools. At the end of each years of instruction or semester and now days it conducted in our country at the end of each term of educational course. Its purpose is to specify students condensed learning and the amount of goals achievement. The result of this method is used to solve students problem them to the higher levels and revise the educational method.[6]

Conclusion:

The educational system with continues and formative assessment are much more successful than others. If there is no evaluation in the educational system, its hard to distinguish between success and failure and when the success is not distinguished benefits can not be devoted to it. If the failure is not specified it con not be solved. If the result of educational system are to be presented by the evaluation measure, appropriated can be attained. One requirement should be added and this is evaluation which must be right and arranged. Right now Iranian educational system has taken a lot of flak. Unfortunately in our educational system the course content is the only criteria which is to be taken into account. Our educational assessment system is traditional one which discuss only the content of the books and does not related to our world out of the books. The students evaluation must be indispensable part of teaching and learning process and not the and, in the evaluation system. Group activity must be assessed

as well as individual activity. Regarding the different purposes kinds and characteristic of evaluation, different methods and measures of evaluation such as oral questions, practical exams, submitting articles, practical evaluation, self assessment are used. In the process of teaching and learning teachers should take factors like purposes and content into account and according to that they should choose their appropriate methods-like formative, summative, collective, combined, national and so on.

Reference:

- [1]. A.Kiamanesh “assessment and measurement in the science and mathematic“,Knowledge Publications,,vol 2,(2009),p.111-117
- [2]. B.Behzadian ”criticism on the educational institutes evaluation” ,University Press ,vol2,(2010),p.129-131
- [3]. A.seif “methods of educational measurement and evaluation” ,Research Institute, Tehran ,vol2,(2009),43-50
- [4]. H.Ahmadi, and M.Hasani ” New descriptive criterion evaluation in the educational assessment”, Sadra Publications ,vol3, (2011),p.77-81
- [5]. T.Rastgar ”Evaluation at education service” , Science Publications,vol4, (2010),p.13-19
- [6]. M.Farahani ”Introduction to qualitative evaluation of learners learning”, Publishing Scholar ,vol1,(2010),p.2-9

Examining characteristics, obstacles, reasons and process of smart schools' establishment

Mehdi H.Soltani, Alovzat Q.Aliyev
Institute of Information Technology of ANAS, Baku, Azerbaijan
soltani466@yahoo.com; alovsat@iit.ab.az

Abstract

By improvement of information technology in world ,in educational systems this technology is the best .In modern schools using new technology caused improvement in students education. Using of smart schools is the most advanced kind of these schools. In this essay we will examine important characteristics of these schools that cause leading to smart schools and establishment reasons of smart school , obstacles and problem that are in establishment of these schools in educational system , process of change that should established in educational structures such as students, teachers, managers and even parents.

Keyword

computer; establishment ; smart; Schools; process

Introduction

Common education in country's schools is a traditional education. In other words, it is just listening. The new method will be tried to presented education with training films, animation, and etc. In this way, stability of subjects is more than traditional way. Also in the method, scientific subjects can be transmitted in a more attractive environment and less time.

In traditional schools, lesson plan of teacher includes a set of guidelines, curricula, training additional questions, exams and class ... Be. But in addition to these new schools, teacher of multimedia educational materials including video, photos, audio, slide and ... Uses the quality and shelf life to promote education. That is the first step on the way toward smart schools. Of course, if we imagine New information technology (in its different aspects: Internet, computers, multimedia and ...) makes the educational revolution lonely, because if the training-learning culture in the educational system does not transform, New information technologies will not only bring changes, but to strengthen the conservative traditions will lead the training. Therefore, traditional schools moving toward smart school requires a change in the country's educational system and never smart schools will not be possible unless the structure of education system be changed. This requires long

term planning. This change should be step by step and with tact and thought. [1]

The most important features and characteristics of smart schools:

A) In smart Schools, teachers can design new courses according to the needs and interests of students with using of databases and software programs, etc or modify and improve existing courses .So the educational content of courses in these schools will be different from other schools.

B) Students of schools determine their own learning rate. Also in these schools, hours learning are not limited to school hours and when students will their favorite classes, they will be provided through computer applications or telecommunications in every time.

C) The role of teachers in these schools, will be changed largely from the practicum students to pursue their private education and therefore they will find more leisure and opportunity for their professional growth and development (reading, communication and constructive interaction and effectively with colleagues, promote and improve science teaching skills and ...).

D) In these schools, students often present in the classroom rather than large bags full of books with computers (lab top). In other words, in addition to current teaching materials and textbooks, in these schools, various software and multimedia software courses such as CD and ... Shall be used. [2]

E) In these schools, Evaluation of students is done every day and continuously, rather than in indented sections shifts (at the end of each chapter or each term and ...), and also some of these tests are done online and remote.

F) In these schools, when students enter and exit the school, with digital cards that they own them, their parents are made known of enter/exit times (SMS sending through school automation system). Parents can connect to the school evaluation system with a simply training and take information about status of self-child's academic and gain knowledge progress. In These school, an E-Mail usually will be sent to

parents at the end of class hours from the school for them to inform of children state in that day.

G) Since some of the traditional activities in other schools are decreased by information technology in these schools, but quality and quantity of interactions between students, teachers and parents will promote and because of result from interactions of these three groups, Possibility of creating better learning conditions for students is provided. [3]

I) information technology is the exploitation technique of human ideas that concedes known, repetitive and non-creative works to machine through automated operations and makes free the human thoughts toward the unknown revelation. Peter Drucker said “handwork give its place to knowledge work, but determinant role of humans will continue and establish as the organization dominant”.

So using of these technologies is only as a tool for achieving the goal of education and more efficient training and appropriate with incremental changes in the world.

The most important reasons of establishing Smart schools:

A) Nowadays, due to the growth of computer technology, speed of data transfer and explosive of knowledge, information and knowledge can be available to everyone easily and quickly. In such circumstances, using of information and informatics technologies in smart schools provides the updating scientific information and enhancing teaching skills of teachers so that they can estimate more accurate of students’ conventional knowledge using of available tools in these schools and according to students’ conventional knowledge, they coordinate training courses and materials of courses with conventional knowledge of their students. [4]

B) In the other hand, educational programs in traditional schools are mostly based on teacher and are not appropriate with talents, abilities, needs and learning styles of students that each of them has its own rhythm. Smart schools can be useful and effective in order to eliminate or reduce this educational gap due to the flexible curriculum, the possibility of teaching new techniques, having a wide range of programs and teaching methods and giving importance to student role (with considering Individual differences and more attention to their needs and interests and talents). In fact each student can be trained depending on his/her talent and in other words, education system is variable to students’ potential.

C) The future information society requires people that are able to use information technology for growth and development. In this age, remain deprived of day knowledge, insight and skills leads to unemployment, social inequalities and thus the emergence of dissatisfaction and stress. Smart

schools are planned mainly to supply these needs because in these schools students learn how extract their need information through information networks, how think about them and how use their findings in order to solve self problems and development of their communities. [5]

The existing barriers in smart schools’ development way:

– It disarrange the existing order and has concerns with itself. It provokes a negative kind of opposition and resistance in humans. (Especially concerns are introduced such as the possibility of failing students in the school entrance, and possibility emotional problems and Heart scurry in students to a new school system and education methods, etc.)

- Inadequate skilled and trained manpower.

- Lack of financial and physical resources to equip smart schools.

- Lack of proper vision of information technology among the people and officials, especially teachers and administrators and the education and ability to apply the tools and resources, environmental conditions and public opinion understanding of this phenomenon is more important than all. [6]

- Thought to jeopardize job security of teachers with innovations

- Economic problems of many families to provide at least one computer system

- Lack of true cultural beliefs of our society, especially parents of students about smart schools

The process of creating change and achieving to smart schools:

- At first we must have the deep understanding and thought of these schools.[7]

– we must begin from universities and teacher training centers, managers and all education personnel. These people should learn how as students in educational environments to learn and grow, and try to flourish in the chase to the unique talents of students, have professionals interaction and cooperation tend, pay to assess their performance, have an active presence in determining Educational Policy and be able, to create a dynamic environment in the area of education that changes with changing educational needs of society. [8]

- A serious transformation in training materials must be done, and educational materials suppliers should be responsible for this work. Best ways of thinking with course topics should be

blended and also the structure and relationship between different topics to be determined accurately. [9]

- Perform activities that students have incentive and facilities to do them and normally they will lead to optimal learning. Therefore, educational materials for the students to learn well and what they learned to take properly, you must create the appropriate models, promotion of thinking and create motivate in students, a bridge shall establish between what they have learned and the new position.

- School and learning environment with the efficient using of existing technologies to provide and strengthen the above methods, try to upgrade imagination of student to their abilities and capabilities.

- Evaluating is a tool in the best position to reflect students' learned and it must create a dynamic environment in which student can be apply his/her findings and learned in solving problems that are new, but they are according to his/her skills, or They are dealing with them in everyday life or interested in -solving them. In such an environment student has the greatest sense of responsibility to quality of learning and doing things.

- Evaluation is usually focused and is done just with clear and precise criterions without qualitative and quantitative comparison of students works and in addition to teachers, it is available for the students also. However, it is necessary to evaluate the performance of teachers.[10]

Conclusion

Mr. David Perkins, one of the smart school designers, said that the smart school has two basic ideas:

1. Learning is result of thinking and all students can learn to think properly.
2. Active and appropriate use of knowledge, will lead to deep understanding and learning should be included deep understanding.

Most important characteristics of today are the growing developments in momentum of scientific, technological, social and... . When only stable phenomenon is change and instability, the of human communities and organizations are inevitable of access to new trends for survival , dynamic and creating constructive developments in the future, because according to Toffler said "Only with innovative using of change is for its conductivity changes that we can be spared future shock damage to a better future and achieve humane. On the other hand, in all societies ,the education institutions are expected to be the head is the source of social changes and innovations in addition to recreate in culture and transmission to the next generation , because the education system creates major infrastructure of personalities and social perspectives

based on its mission, an then if it does a conventional collaborative effort , expected innovation would be easier in the community. This means that education system should also be able to coordinate their development towards modern society and predict the future changes, in order to lead changes to create favorable developments in the future.

In this regard, one of the approaches that can be helpful in responding to these needs in the educational systems and today in many developed countries implemented or are being implemented, the expanded using of ICT in education systems and consequently the establishment and development of smart schools. actually These schools was excluded invention from a kind of philosophy and new concept of education, that they try applying information technology ,with Elimination a series of traditional deterrence, to present training. Effective education in such schools requires that students get new roles in learning, therefore that is information seekers and are able to arbitrate and evaluate the value of the extensive data on the internet that they are available for using. In such circumstances, the role of teachers is changed as well as transfer knowledge and information to facilitate the learning process. They should also work in such a way that students can take self-confidence, information management strategies and necessary skills to achieve negation in order to everyday life and they be able to use them in their working environment and technological tools of mass communication successfully.

As was mentioned, like any other educational innovation in the education system, there are obstacles to the establishment and development of these schools, that the major of them are cultural and structural problems. in the last 50 years, Many changes has happened in human life, if you look around your life , you see those, but unfortunately classrooms has not changed than 50 years ago and training is done with Blackboard. If the officials and the administrators, teachers and people do not accept that this traditional style replications is completed ,then they will never accept the new methods, because infrastructure "change" and "change by" is readiness and interest to accept it, otherwise change does not happen necessarily and does not have predicted results and always associated with resistance and defensive of individuals. we should scrutinize Another important point in this regard . it is that we should know technology is a tool and efficient application of any technology kinds comes from thought, culture and social relations deeply. Therefore, we require improved approaches, review educational policies, reorganize the content, improve human resources, effective curriculum design and development of cultural criteria to provide coexistence with the new technology for Effective and efficient utilization of information technology, particularly in developing education.

References:

- [1] A.Toffler “Future Shock” interpreter: Heshmatollah Kamrani , Publications of the Institute of Education Research Institute, Tehran,vol. 2, (2009)
- [2] J.Thomas “Global Issues of Education” interpreter, Ahmad Aghazadeh, vol. 2, (2008)
- [3] M.Saatchi “Applied Psychology for Managers”(2009)
- [4] A.Solimani “six primary technological learning in communities,technology, Internet”, Sun Publications stamped, vol. 1, (2010)
- [5] A.Majidi “Superior system” vol. 1, (2010)
- [6] F.Mashayekhi”New perspectives in educational planning” Publications of the Institute of Education vol. 2, (2007).
- [7] M. Griffin “organizational behavior“ vol. 1, (2010)
- [8] M.Soroush” Expert IT communication” vol. 1, (2010)
- [9] M.Niknami” supervising educational” vol. 2, (2009)
- [10] M.Tusi “Journal of Management in Education”,vol2, (2009)

Factor Analysis of Factors Affecting E-Learning Success from the Viewpoint of Virtual Students (Case Study of Islamic Republic of Iran)

Aref Riahi, MA in Information Science, University of Tehran, Iran Corresponding Author, Email ; Ariahi@ut.ac.ir ,
Call Number +981525264700

Hasan Khosravi, MA in Educational Management, University of Tarbiat Moaalem, Tehran, Iran

Samaneh Khakmardan, MA in Publishing Management, University of Imam Reza, Mashhad, Iran

Abstract

Recognition and familiarity with factors influencing the effectiveness of E-learning for avoiding mistakes and subsequently failure seems necessary. Considering this necessity, the purpose behind this research is recognition and categorization of factors affecting E-learning success in Iran's higher education. The researcher used library methods and field research in order to attain the research's objective. Present research's statistical population includes virtual students of Shiraz, Khajeh Nasir, and Sahand university of Tabriz. The sample size is determined through utilization of Morgan table. This research's main tool was a questionnaire designed electronically in web after validity and reliability confirmation. Data are processed and analyzed through utilization of SPSS. Results have shown that "selection of appropriate educational media", "interaction possibility in educational media", "educational content", and "appropriate software utilization" are among factors influencing E-learning success. Factor analysis of factors affecting E-learning success has shown that these factors can be divided into two categories of support factors and content, as well, educational tools.

Keywords: E-learning, Virtual Student, Factor Analysis, Iran

Introduction

Global developments process is developing with a focus on IT phenomenon development. Learning process has been changed and transformed concurrent with rapid changes in techniques and skills and the emergence of new phenomenon in IT and their impacts on methods and ways of living. Existence of extensive communication networks, especially internet and advanced educational tools, has changed educational methods; besides, it has provided a situation for a large population of knowledge seekers across different parts of the world and far distances get covered by distance learning network and trained by methods different from traditional and usual training methods [1]. Distance learning methods have been developed along with other sciences and technologies; nowadays they have been developed to the extent of utilizing different kinds of technology and products such as computer, communication, and digital networking technologies and lead to the formation of issues such as E-education and E-learning.

E-learning has been defined as a system based on technology, organization, and management. E learning gives students the necessary ability through internet and facilitates their learning in this process[2]

. E-learning has evolved as a new pattern in modern education; besides, it has a large influence on schools, institutes, and scientific and research organizations, though its major effect has been on universities and higher education centers [3] & [4]

Theoretically, E-learning can bring about many benefits. Students have a high flexibility for learning in E-learning. They can learn

each time and each place as they wish. In most cases, learners are satisfied with what E-learning presents at a risk-free environment in both academic and corporate collections. Besides, in E-learning, especially the kinds of education presented in educational centers, education time has been less and education costs have been decreased 50% to 70 % [5].

Furthermore, many researchers and reports, published throughout the world, emphasize many times that there is not much difference between students' educational achievements through E-learning and face-to-face learning; though this kind of learning is at a lower level of quality in comparison with face-to-face learning in a specific space and time. Therefore, E-learning programs should render appropriate qualitative indices; furthermore, this learning's skills and capabilities should be introduced more and more. Each university or institute intending to implement E-learning should be familiar with the factors affecting their success in order to decrease the risk of defeat. The present research's main objective is the recognition of factors having a major role in E-learning course and providing success in E-learning process.

Research Background

Many researches and studies done on E-learning and the factors affecting its success, we will mention them very briefly. Popp (2000) introduces those factors affecting E-learning success as following: intellectual property, the appropriateness of the period for E-learning environment, maintenance and measurement of an electronic period's success [6]. The results of the study done on the quality of Australia higher education's E-learning have shown those factors affecting E- Learning success: trainer's expertise including teaching on-line, using technology for teaching, trainers' training; students' readiness including technological skills, access to technology, technology literacy, self-regulated learning; technology infrastructure including: E-learning systems, hardware and software, provision of content services and necessary resources for learning, training plans [7]. Govindasamy mentions seven major factors affecting E-learning success: organizational support, content codification, teaching and learning, course structure, supporting students, supporting scientific board members, and evaluation [8]. Chiu et al consider these three factors, usefulness, quality, and observed value in learning system, affecting the satisfaction of E-learners [9]. Khan also categorizes those factors affecting E-learning in eight categories: educational factors, technology, interface design, management, resources support, human factors, organization factors, and evaluation [10]. Frazzen mentioned in his thesis "factors affecting learning quality with web support" the relation and effect of several major factors that are: organization factors, educational factors, trainer, students,

technology, and educational design [11]. Pei-Chen Sun et al. consider trainer's anxiety about computer, teacher's viewpoint of E-learning, flexibility of E-learning period, quality of E-learning period, usefulness, facilitation of usage, and variety of evaluation among factors affecting learners' satisfaction of E-learning periods. Jen-Her, Tennyson and Hsia mention factors such as adequacy of working on computer, system performance, content quality, interaction, executive expectations, and learning climate among major indices affecting the student's satisfaction in a combinational E-learning system [12].

Methodology

The researcher first identified and extracted the factors affecting the educational quality of E-learning system through method. In the second step, he prepared a questionnaire considering research purpose and the results obtained from library study. Study society includes virtual students of Shiraz, Khajeh Nasir, and Sahand university of Tabriz in computer engineering, and IT engineering. The total number of virtual students in these majors has been 455. Random sampling has been used for determination of sample size; 200 students are selected and questionnaires are sent to them electronically. SPSS is used for data analysis.

Findings

Several factors are intervened in the success of educational systems especially E-learning system. A list of factors affecting E-learning system success has been extracted considering the results obtained from resources and studies review. Likert Spectrum is also used to measure the importance of each factor mentioned above.

Table 1 .Factors affecting E-learning success from the viewpoint of virtual students

	Factors	Mean	SD
1	selection of appropriate educational media	4.82	0.42
2	The possibility of interaction in educational system	4.80	0.41
3	Educational content	4.75	0.51
4	Utilization of appropriate software	4.74	0.49
5	Easy access to educational materials	4.69	0.57
6	Research incentives for scientific board members	4.64	0.56
7	Information and Communication Technology infrastructures	4.55	0.49
8	Helping students to solve their	4.54	0.61

	problems during a period		
9	Utilization of new learning strategies	4.51	0.58
10	scientific board members' service training	4.50	0.77
11	Training needs assessment	4.47	0.55
12	Appropriate organization of educational materials	4.47	0.68
13	Interaction rate among learners	4.43	0.71
14	Usual daily student services to virtual students	4.38	0.76
15	Appropriate information management	4.29	0.81
16	Holding necessary additional courses	4.22	0.95
17	Cooperation of those involved in education system	4.19	0.88
18	Reduction of scientific board's required units	4.16	0.92
19	Team assignments	4.09	0.77
20	Supporting scientific board members	4.02	1.05
21	Exercising the intellectual property's rights	4.00	0.94
22	Providing technical consultation about the utilization of electronic system	3.97	1.11
23	Appropriate management of registration and admission.	3.89	1.12

Data of the above table indicates that all the issues investigated, other than two issues relating virtual students' viewpoint, have an average above 4 (very important and important). Among those factors, the most important ones are: selection of appropriate educational media and possibility of interaction in education system.

Factor analysis is used to categorize those factors affecting educational quality of E-learning from the viewpoint of virtual students. KMO coefficient and Bartlett *statistic* used to determine the appropriateness of collected data for factor analysis. In this section, KMO equals 0/921 which indicates the appropriateness of existent correlations between data for factor analysis. On the other hand, Bartle Test is used to assure the appropriateness of data for factor analysis. Bartlett statistic obtained 5122/31 that was meaningful at 1% level. Therefore, data were appropriate for factor

analysis. Number of extracted factors, along with eigenvalue of each, variance percentage of each factor and variance cumulative percentage are shown in table2.

Table 2. Summary of factor analysis of those factors affecting education system and E-learning

Factors	Eigenvalue	Eigenvalue variance percentage	Variance cumulative percentage.
Factor 1	11.920	48.907	48.907
Factor 2	5.558	23.382	72.289

As table2 shows, these two factors could determine 72/29 percents of the total variance of factors affecting E-learning system success. Varimax method is used for factor rotation. After rotation stage, those variables related to each factor are clarified in columns. After processing representative variables of factors affecting education system and E-learning, those obtained factors are named.

Table3 shows each factor and those variables related to that factor along with load factor and the buoy lining. Factor analysis results have shown that interactions and support factors have assigned 49/71 percents of variance to themselves and been the first success factor of E-learning system with 11/93 eigenvalue. Furthermore, these results have indicated that educational content and tools have allocated 23/28 percents of variance to themselves.

Table3. Factors variables and load factor resulted from rotation matrix

factor name	variables	Factor loading
fator1: interactions and support factors	interaction rate among learners	0.859
	cooperation rate among those involved in education system	0.843
	team assignments	0.761
	Holding training courses for scientific board members	0.806
	Preparation for launching virtual courses	0.903
	Reduction of required units for scientific board members	0.602
	Possibility of education system for providing interaction	0.740
	Supporting scientific board members during the period	0.820

factor2: educational content and tools	Training needs assessment before periods initiation	0.880
	Selection of appropriate educational media	0.624
	Utilization of appropriate software	0.629
	Appropriate organization of educational materials	0.836
	Appropriate educational content	0.795
	Utilization of new learning strategies".	0.716

Conclusion

E-learning is a new paradigm and the product of IT; it leads human beings towards a major revolution in education. E-learning is a key for transition of manpower into information society. In the present research, those factors affecting E-learning system success from the viewpoint of virtual students has been investigated. Factor analysis of factors affecting E-learning system success has shown that those factors can be categorized into two categories of support and content factors and educational tools. In the present research, interactions have been divided into two factors of team tools and individual tools through using factor analysis; besides, team tools have a higher priority. Following suggestions are offered on the basis of findings, also for increase in education quality of E-learning system:

Possibility of usage of methods and interactive teaching techniques through enhancing capabilities of learning management system (LMS).

Scientific board members of virtual course should use those methods and teaching techniques helping the increase in educational interactions among trainer and learners and those interactions among learners themselves

considering the importance and role of scientific board members on E-learning system effectiveness, those scientific board members should be used in this system which have characteristics such as ability for management and learners' encouragement, virtual presence and interaction in education process, supporting students,

E-commitment and skill, capability for providing an interactive environment, and having a positive attitude.

References

- [1]. Gulati, S. Technology-enhance learning in developing nations: A review. *International Review of Research in Open and Distance Learning* 9(1) : 1-16, 2008.
- [2]. Levy, Y. Assessing the value of E-learning systems. USA: Infosci, 2006.
- [3]. Pei-Chen, Sun, Ray J. Tsai, G. Finger, Yuch-Yng Chen and D. Yeh. What drives a successful E-learning? An empirical investigation of the critical factors influencing learner satisfaction. *Computers & Education* 50 (1): 222-714, 2008.
- [4]. Harper, K.C., K. Chen, D.C. and Yen. Distance learning, virtual classrooms, and teaching pedagogy in the internet environment. *Technology in Society* 26(4): 585-598, 2004.
- [5]. Turban, Efraim and et al. *Information Technology for Management: Transforming Organizations in the Digital Economy*. USA: Wiley, 2006.
- [6]. Papp, R. Critical success factors for distance learning. USA: American conference on Information Systems, Long Beach, 2000.
- [7]. Oliver, R. et al. Flexible toolboxes: a solution for developing online resources, in Lockwood, F. and Gooley, (Ed) *Innovation in open and distance learning*. London, Kogan Page, 2001.
- [8]. Govindasamy, T. Successful implementation of E-learning pedagogical considerations. *The internet and higher education* (4) 287-299, 2002.
- [9]. Chiu, C.D. M. H. Hsu, S. Y. Sun, T. C. Lin, and P. C. Sun. Usability quality, Value and E-learning continuance decisions. *Computers & education* 45 (4), 399-416, 2005.
- [10]. Khan, B.H. *Managing E-learning: Design delivery, implementation and evaluation*. Hershey, PA: information scientific publishing, 2005.
- [11]. Frazeen, B. Technology to enhance the learning experience. Available at: www.clomedia.com/contact/templates/clo_feature.asp?articleid=218, 2006.
- [12]. Jan-Her, W., R.D. Tennyson, and T. L. Hisa. A study of student satisfaction in a blended E-Learning system environment. *Computer & Education* 55 (1) : 1-39, 2010.

I Did Not Know Its Prohibited - Academic Dishonesty in Online Courses

Yovav Eshet

Yehuda Peled

Keren Grinautski

University of Haifa & Zefat Academic College, Israel

E-mail: YovavE@Zefat.ac.il

Abstract- There is a disagreement among researchers in regard to academic dishonesty in online as compared to traditional learning settings. Based on this, the aim of the current study was to investigate the connection between academic dishonesty in the virtual versus face-to face teaching/learning settings in relation to students' learning motivation, while examining the phenomenon from a cross-cultural perspective. The sample consisted of 1,574 participants - 803 from USA and 771 from Israel. The results showed that there are significant differences in students' likelihood to engage in academic dishonesty based on the type of course, such that students in face-to-face courses are more likely to engage in acts of academic dishonesty than their counterparts in online courses. In addition, it was found that students' propensity to engage in academic dishonesty is explained by motivational orientation, type of course, and age. The findings were consistent across student groups in both countries. The phenomenon can be explained by the fact that more intrinsically motivated students self-select online as opposed to traditional classroom courses, and because online instruction facilitates increasing levels of intrinsic motivation.

I. Introduction

With the rapid growth of distance learning involving the Internet, there is a greater opportunity for individuals to engage in Academic dishonesty and plagiarism, particularly where there is little or no personal contact between students and faculty [1], [2]. Kelley and Bonner [3] suggested that students who feel close to their professors tend to be more honest. However, the ability for faculty to develop a strong rapport with students becomes more difficult in the online learning environment. Students who feel "distant" from others are more likely to engage in deceptive behaviors, such as cheating [4], [5]. Online courses, as contrasted with traditional classroom courses, may serve to exacerbate these feelings of separation and thus, may contribute to the incidence of academic dishonesty [6] [7] [8]. Both students and faculty perceive that cheating is more likely to

occur in online rather than face-to-face classrooms [9] [10].

Conversely, there are some reports suggesting there is less academic dishonesty in online as compared to traditional learning settings. The reason for this latter finding is that academic dishonesty may be associated with the extrinsic motivation that drives students in traditional courses [11]. Online students may be more intrinsically motivated by being able to learn independent of traditional classroom settings, while this type of motivation could substantially reduce their desire to cheat [10].

II. Motivation and E-learning

According to Deci and Ryan [12] there are two types of motivation: intrinsic and extrinsic, which are based on different reasons or goals underlying an action. *Intrinsic motivation* refers to doing something because it is inherently interesting or enjoyable, while *extrinsic motivation* refers to doing something because it leads to an enjoyable but external and separable outcome [13]. Self-determined motivation was found to be related to more interest, effort, positive emotions, satisfaction, and commitment by students.

Intrinsic and extrinsic motivations can be viewed as extremes on a continuum with additional types of motivation that vary according to level of self-determination [14] [13]. Those are (1) *intrinsic motivation* explained above; (2) *regulation through identification* -this is the most autonomous or self-determined form of extrinsic motivation. Here, a person considers his activity as important and beneficial and he carries it out, although he does not enjoy it; (3) *introjected regulation* - individuals begin to internalize the reasons for their actions, but they do things in order to avoid feelings of guilt or anxiety or to attain ego-enhancements or pride [15];(4) *external regulation* behaviors are performed to satisfy an external demand or to obtain an externally imposed reward contingency [16]; and, finally, *amotivation*, which is external to the motivation continuum, refers to a state where intention to act appears to be absent.

Motivation plays an important role when one chooses to participate in an online course [17] [18] as intrinsic motivation is considered to be a significant predictor of persistence and achievement in distance

education [19] [20]. In addition, Grolnick and Ryan [21] found that controlling environments reduce a student's sense of autonomy, decrease intrinsic motivation, and result in poorer performance.

George and Carlson [22] contend that as the distance between a student and a physical classroom setting increases, so does the frequency of cheating. Their assumption, coupled with the belief that academic misconduct is more pervasive in the virtual classroom [9] [6] [7] [23], led us to question whether there is (1) a higher incidence of cheating in online courses as compared to traditional on-campus face-to-face courses; and (2) a relation between student's motivation, type of course, and frequency of academic misconduct. Thus, the purpose of this research is to explore the connection between self-reported frequencies of academic dishonesty in the virtual versus face-to-face teaching settings and students' learning motivation. Another purpose is to examine, for the first time, the issue of academic dishonesty from a cross-cultural perspective, i.e., by comparing online and face-to-face students in U.S. and Israeli academic institutions. We hypothesize that Israeli students will report less cheating than their counterparts in USA. This is due to the fact that Israel is high-uncertainty avoidance culture [24] in which people are not likely to engage in deviant behavior [25]. Conversely, people in low-uncertainty avoidance cultures, such as the United States [24], may engage in deviant behavior, since they tolerate risks easily [26]. Based on the foregoing, we hypothesize that there will be differences in the level of motivation between students that learn in traditional settings and e-learners in motivation and propensity toward academic dishonesty. Specifically, e-learners will show higher levels of intrinsic motivation and less propensity toward academic dishonesty than learners in traditional face-to-face settings.

Method

A. Participants

The sample consists of 1,574 participants with 803 from two American academic institutes and 771 from four Israeli academic institutes. 65% of the participants were women and 35% were men. The age ranged from 17 to 59 (mean is 26.4 years). 26% of the participants were freshmen, 32% - sophomores, 20% - juniors, 19% - seniors, and 3% were graduate students. 46% were Christians, 38% were Jews, and 16% were Muslims. 13% of the participants were excluded from the analysis because their surveys were incomplete or carelessly completed. Therefore, the final data set consisted of 1,376 participants.

B. Survey Instrument

A three part survey instrument: part 1 contained 16 items compiled from the Academic Self-Regulation Questionnaire (SRQ-A) [14]. The questionnaire examines four types of motivation: external regulation, introjected regulation, identified

regulation, and intrinsic motivation. The reliability of this questionnaire, measured by Cronbach's alpha 0.79

Part 2 contained questions that examined academic integrity using the Academic Integrity Inventory [27]. These questions inquired students' likelihood to engage in various forms of academic misconduct. The instrument was validated by [27] and the reliability of this questionnaire, measured by Cronbach's alpha, was 0.75. Part 3 contained a series of socio-demographic questions.

C. Procedure

In order to encourage the participants to think in the frame of a specific type of course, we administered a printed version of the survey instrument in the traditional face-to-face courses and an on-line version of the survey instrument in the e-learning courses. The survey instruments were coded and grouped according to the location of the participants' college or university (USA or Israel).

D. Results

Table 1 summarizes the results of Independent Sample T-test analyses, which indicate that there were significant differences in the level of motivation between students attending face-to-face or e-learning courses. These differences were found for three of the four motivational orientations (introjected, identified, and intrinsic) in the U.S. and for all four orientations (extrinsic, introjected, identified, and intrinsic) in Israel. The findings are presented according to the location of the participants' college or university.

TABLE I

Differences in motivational orientation by course type and country

Country	Motivation type	Course type	N	Mean	S.D.	T-Test
USA	Extrinsic	E-learning	287	2.61	0.65	0.02
		Face-to-Face	476	2.61	0.62	
	Introjected	E-learning	287	3.23	0.60	2.72**
		Face-to-Face	477	3.11	0.56	
	Identified	E-learning	287	3.77	0.42	4.82***
		Face-to-Face	477	3.61	0.51	
	Intrinsic	E-learning	287	2.82	0.62	9.03***
		Face-to-Face	475	2.37	0.74	
Israel	Extrinsic	E-learning	293	2.37	0.61	2.13*
		Face-to-Face	316	2.48	0.65	
	Introjected	E-learning	293	2.88	0.61	15.50***
		Face-to-Face	316	2.13	0.57	
	Identified	E-learning	293	3.61	0.54	44.60***
		Face-to-Face	318	1.53	0.61	
	Intrinsic	E-learning	293	2.85	0.66	9.78***
		Face-to-Face	316	2.31	0.70	
Overall Sample	Extrinsic	E-learning	580	2.49	0.64	1.95*
		Face-to-Face	792	2.56	0.64	
	Introjected	E-learning	580	3.05	0.63	8.85***
		Face-to-Face	793	2.72	0.74	
	Identified	E-learning	580	3.69	0.49	19.92***
		Face-to-Face	795	2.78	1.16	
	Intrinsic	E-learning	580	2.84	0.64	13.24***
		Face-to-Face	791	2.35	0.73	

***P<0.001, **P<0.01, *P<0.05

The data in Table 1 indicate that, students in e-learning courses had significantly higher levels of intrinsic motivation than those in face-to-face courses.

In the overall sample, there was a statistically significant difference in the level of extrinsic motivation with students in e-learning courses having lower levels as compared to students in face-to-face courses. This difference in level of extrinsic motivation was not found for students in the U.S. In general, no significant differences were found between American and Israeli students in the levels of academic dishonesty.

TABLE II

Differences in academic dishonesty by course type

Country	Course type	N	Mean	S.D.	T-Test	F
USA	E-learning	287	2.03	0.83	10.334***	24.351***
	Face-to-Face	468	2.73	0.99		
Israel	E-learning	291	2.33	0.95	2.601*	
	Face-to-Face	311	2.52	0.86		

***P<0.001, **P<0.01, *P<0.05

Table 2 summarizes the results of Independent Sample T-test analyses, which indicate that there were statistically significant differences in students' likelihood to engage in academic dishonesty based on the type of course in which they were enrolled. Specifically, it was found that students in face-to-face courses were more likely to engage in acts of academic dishonesty than their counterparts in e-learning courses. Based on MANOVA analysis we found significant interaction between country and course type ($F_{(1, 1353)} = 24.35, p < 0.001$).

TABLE III

Stepwise Regression analysis – motivational orientation, type of course, and socio-demographic variables as predictors of academic dishonesty

	Predictors	β	t	F	R ²	R ² A
Step I	Country (0=USA, 1=Israel)	-0.08	2.46*	18.71	0.11	==
	Teaching Method (0=E-learning, 1=Face-to-face)	0.35	7.19***			
	Gender (0=Female, 1=Male)	0.01	0.53			
	Age	-0.10	3.20**			
	Course type (0=Optional, 1=Required)	0.03	0.82			
	Average Grade	-0.08	1.95			
Step II	Country (0=USA, 1=Israel)	-0.10	1.43	12.49	0.12	0.01
	Teaching Method (0=E-learning, 1=Face-to-face)	0.35	7.00***			
	Gender (0=Female, 1=Male)	0.01	0.36			
	Age	-0.09	2.91**			
	Course type (0=Optional, 1=Required)	0.02	0.78			
	Average Grade	-0.07	1.74			
	Extrinsic Motivation	0.12	3.37**			
	Introjected Regulation	-0.08	1.59			
	Identified Regulation	0.03	0.40			
	Intrinsic Motivation	0.00	0.18			

Table 3 summarizes the results of a Stepwise Regression analysis used to explain the effect of

motivational orientation on academic dishonesty. Likelihood to engage in acts of academic dishonesty served as the dependent variable and motivational orientation along with socio-demographic factors served as the independent variables. The hypothesis that students' propensity to commit acts of academic misconduct would be related to type of course (face-to-face versus on-line) and motivational orientation was supported. In addition, there is a correlation between motivational orientation and academic misconduct ($r_p = 0.115, p < 0.001$), however, when controlling for course type using Partial Correlation Analysis the correlation coefficient is reduced ($r_p = 0.075, p < 0.01$), confirming that course type mediates the relationship between motivational orientation and academic misconduct.

The results of the regression analysis indicate that approximately 13% of the variance in students' propensity to engage in academic dishonesty is explained by motivational orientation, type of course, and age. Specifically, students' likelihood to engage in dishonest acts was found to vary directly with the level of extrinsic motivation and participation in face-to-face courses (as opposed to e-learning courses) and inversely with age. Simply put, the regression results show that the only motivational orientation found to explain students' likelihood to engage in academic dishonesty was the extrinsic motivation. More specifically, the more students are extrinsically motivated, the more likely they are to engage in academic dishonesty. Course type was also found to explain academic dishonesty. According to the results, students in face-to-face courses were more prone to engage in academic dishonesty than e-learners. Finally, age of students was found to explain academic dishonesty, with younger students inclined to cheat more than older students.

Discussion and conclusion

The results of this study are in accordance with the findings of [11], who suggested that academic dishonesty can be explained by extrinsic motivation. Similarly, our findings indicate that of the four types of motivational orientations, extrinsic motivation is the only type that explains academic dishonesty in sample populations of American and Israeli students.

Furthermore, [10] found that there is less overall cheating in the virtual than in traditional classroom settings. They explained that these students may have a higher motivation to learn or able to learn independent of the structure typical in traditional classroom settings, which could substantially reduce their desire to cheat. Our study found that e-learning students manifest significantly higher levels of intrinsic motivation and significantly lower levels of extrinsic motivation than traditional classroom students. Consistent with [10], we also found that e-learners were less likely to engage in acts of academic dishonesty as compared to face-to-face learners – a finding most likely related to their being

more intrinsically and less extrinsically motivated in their course work.

One possible explanation for these results is that more intrinsically motivated students self-select online as opposed to traditional classroom courses. Since less than 6% of higher education students are enrolled in online courses, they most likely are innovators and early adopters who, according to [28] Diffusion of Innovations Theory, may be more internally motivated by factors such as intellectual curiosity.

Another possible explanation for the higher levels of intrinsic motivation observed in e-learning students as compared to students in face-to-face courses is that online instruction facilitates increasing levels of intrinsic motivation. Zhang's [29] research suggests that the e-learning medium provides a learning environment that "emphasizes intrinsic motivation, self-sponsored curiosity and creative situated learning" (p. 4). This rationale is consistent with Cognitive Evaluation Theory [12], which posits that intrinsic motivation is maximized when individuals feel competent and self-determining in dealing with their environment. [13] pointed out that "interpersonal events and structures (e.g., rewards, communications, feedback) that conduce toward feelings of competence during action can enhance intrinsic motivation for that action, because they allow satisfaction of the basic psychological need for competence" (p. 58).

It is important to note that this research study examined academic dishonesty in e-learning and face-to-face settings in two culturally different countries – the USA and Israel. The findings were consistent across student groups in both countries. As such, these findings should be interpreted as having greater generalizability and not limited by cultural specificity.

References

- [1] C. Robinson-Zañartu, E.D. Peña, V. Cook-Morales, A.M. Peña, R. Afshani, and L. Nguyen, "Academic crime and punishment: Faculty members' perceptions of and responses to plagiarism," *School Psychology Quarterly*, vol. 20(3), pp.318–337, 2005.
- [2] J. Walker, "Measuring plagiarism: Researching what students do, not what they say they do," *Studies in Higher Education*, vol. 35(1), pp. 41–59, 2010.
- [3] K. Kelley, and K. Bonner, "Distance education and academic dishonesty: Faculty and administrator perception and responses," *Journal of Asynchronous Learning Network*, vol. 9, pp. 43-52, 2005.
- [4] J. Burgoon, M. Stoner, J. Bonita, and N. Dunbar, *Trust and Deception in Mediated Communication*, 36th Hawaii International Conference on Systems Sciences, 44a, 2003.
- [5] N. Rowe, "Cheating in online student assessment: Beyond plagiarism," *Online Journal of Distance Learning*, 2004.
- [6] M. Heberling, "Maintaining academic integrity in online education," *Online Journal of Distance Learning Administration*, vol. 5, 2004.
- [7] K. Kennedy, S. Nowak, R. Raghuraman, J. Thomas, and S.F. Davis, "Academic dishonesty and distance learning: Student and faculty views", *College Student Journal*, vol. 34, pp. 309-314, 2000.
- [8] D. Stuber-McEwen, P. Wiseley, C. Masters, A. Smith, and M. Mecum, *Faculty Perceptions versus Students' Self-Reported Frequency of Academic Dishonesty*, Paper presented at the 25th Annual Meeting of the Association for Psychological & Educational Research, Emporia, KS, 2005.
- [9] T. Grijalva, J. Kerkvliet, and C. Nowell, "Academic honesty and online courses," *College Student Journal*, vol. 40, pp. 180-186, 2006.
- [10] D. Stuber-McEwen, P. Wiseley, and S. Hoggatt, *Point, Click, and Cheat: Frequency and Type of Academic Dishonesty in the Virtual Classroom*, 2009.
- [11] E. Greenberger, J. Lessard, C. Chen, and S. Farruggia, "Self-entitled college students: Contributions of personality, parenting, and motivational factors", *Journal of Youth and Adolescence*, vol. 37(10), pp. 1193-1204, 2008.
- [12] E.L. Deci, and R.M. Ryan, *Intrinsic Motivation and Self-Determination in Human Behavior*, New York: Plenum, 1985.
- [13] R.M. Ryan, and E.L. Deci, "Intrinsic and extrinsic motivations: Classic definitions and new directions," *Contemporary Educational Psychology*, vol. 25, pp. 54-67, 2000.
- [14] R.M. Ryan, and J.P. Connell, "Perceived locus of causality and internalization: Examining reasons for acting in two domains," *Journal of Personality and Social Psychology*, vol. 57, pp. 749-761, 1989.
- [15] J.A. Moreno Murcia, D. González-Cutre Coll, and M. Chillón, "Preliminary validation in Spanish of a scale designed to measure motivation in physical education classes: The perceived locus of causality scale," *The Spanish Journal of Psychology*, vol. 12(1), pp. 327-337, 2009.
- [16] R. deCharms, *Personal Causation*, New York: Academic Press, 1968.
- [17] M.G. Moore, and G. Kearsley, *Distance education. A systems view*, Belmont, CA: Wadsworth, 2005.
- [18] A.P. Rovai, M.K. Ponton, M.J. Wighting, and J.D. Baker, "A comparative analysis of student motivation in traditional classroom and E-Learning courses," *International Journal on E-Learning*, vol. 6(3), pp. 413-432, 2007.
- [19] S. Coussement, *Educational telecommunication: Does it work? An attitude study*, (ERIC Document Reproduction Service No. ED391465), 1995.
- [20] N.F. Fjortoft, "Persistence in a distance learning program: A case in pharmaceutical education," *American Journal of Distance Education*, vol. 10(3), pp. 49-59, 1996.
- [21] W.S. Grolnick, and R.M. Ryan, "Autonomy in children's learning: An experimental and individual difference investigation," *Journal of Personality and Social Psychology*, vol. 52(S), pp. 890-898, 1987.
- [22] J. George, and J. Carlson, *Group Support Systems and Deceptive Communication*, Speech presented at International Conference on Systems Sciences, Hawaii, 1999.
- [23] G.G. Smith, D. Ferguson, and M. Caris, "Teaching college courses online vs. face-to-face," *T.H.E. Journal*, vol. 28, pp. 19-26, 2001.
- [24] G. Hofstede, *Geert Hofstede™ Cultural Dimensions*, 2009, Retrieved on November 30th, 2011 from: http://www.geert-hofstede.com/hofstede_israel.shtml
- [25] P.M. Doney and M. Mullen, "Understanding the influence of a national culture on the development of trust," *Academy of Management Review*, vol. 23, pp. 601-620, 1998.
- [26] G. Hofstede, *Cultural's Consequences: International Differences in Work-Related Values*, Beverly Hills, CA: Sages, 1984.
- [27] J.L. Kisamore, T.H. Stone, and I.M. Jawahar, "Academic integrity: The relationship between individual and situational factors on misconduct contemplations," *Journal of Business Ethics*, vol. 75, pp. 381–394, 2007.
- [28] E.M. Rogers, *Diffusion of innovations*, 5th ed., New York: Free Press, 2003.
- [29] P. Zhang, "A case study on technology use in distance learning," *Journal of Research on Computing in Education*, vol. 30(4), pp. 398-420, 1998.

Solving the Quadratic Assignment Problem with the Modified Hybrid PSO Algorithm

Ali Safari Mamaghani

Department of Computer Engineering
Bonab Branch
Islamic Azad University
Bonab, Iran
Ali.Safari.m@gmail.com

Mohammad Reza Meybodi

Department of Computer Engineering and Information
Technology
AmirKabir University of Technology
Tehran, Iran
mmeybodi@aut.ac.ir

Abstract—In this paper a particle swarm optimization algorithm is presented to solve the Quadratic Assignment Problem, which is a NP-Complete problem and is one of the most interesting and challenging combinatorial optimization problems in existence. A heuristic rule, the Smallest Position Value (SPV) rule, is developed to enable the Continuous particle swarm optimization algorithm to be applied to the sequencing problems. So, we use SPV to the QAP problem, which is a discrete problem. A simple but very efficient Hill Climbing method is embedded in the particle swarm optimization algorithm.

We test our hybrid algorithm on some of the benchmark instances of QAPLIB, a well-known library of QAP instances. This algorithm is compared with some strategies to solve the problem. The computational results show that the modified hybrid Particle Swarm algorithm is able to find the optimal and best-known solutions on instances of widely used benchmarks from the QAPLIB. In most of instances, the proposed method outperforms other approaches. Experimental results illustrate the effectiveness of proposed approach on the quadratic assignment problem.

Index Terms—The Quadratic Assignment Problem; Particular Swarm Optimization; Hill Climbing Approach; NP-Complete problem.

I. INTRODUCTION

The Quadratic Assignment Problem (QAP) is one of the classical combinatorial optimization Problems and is widely regarded as one of the most difficult problem in this class. Given a set $N = \{1, 2, 3, \dots, n\}$ and $n \times n$ matrices $F = \{f_{ij}\}$ and $D = \{d_{ij}\}$, and $C = \{c_{ij}\}$, the QAP is to find a permutation ϕ of the set N which minimizes

$$z = \sum_{i=1}^n \sum_{j=1}^n f_{ij} d_{\phi(i)\phi(j)} + \sum_{i=1}^n c_{i\phi(i)}$$

As an application of the QAP, consider the following campus planning problem. On a campus, new facilities are to be erected and the objective is to minimize the total walking distances for students and staffs. Suppose there are n available sites and n facilities to locate. Let d_{kl} denotes the walking distance between the two sites k and l where the new facilities

will be erected. Further, let $f_{i,j}$ denotes the number of people per week who travel between the facilities i and j . Then, the decision problem is to assign facilities to sites so that the walking distance of people is minimized. Each assignment can be mathematically described by a permutation ϕ of $N = \{1, 2, 3, \dots, n\}$ such that $\phi(i) = k$ means that the facility i is assigned to site k . The product $f_{ij} d_{\phi(i)\phi(j)}$ describes the weekly walking distance of people who travel between facilities i and j . Consequently, the problem of minimizing the total walking distance reduces to identifying a permutation ϕ that minimizes the function z defined above. This is an application of the QAP with each $c_{ik} = 0$. In this application, we have assumed that the cost of erecting a facility does not depend upon the site. In case it does, we will denote by $c_{i,k}$ the cost of erecting facility i at site k , and these costs will also play a role in determining the optimal assignment of facilities to sites [1]. Additional applications of QAP include the allocation of plants to candidate locations, layout of plants, backboard wiring problem, design of control panels and typewriter keyboards, balancing turbine runners, ordering of interrelated data on a magnetic tape, processor-to-processor assignment in a distributed processing environment, placement problem in VLSI design, analyzing chemical reactions for organic compounds, and ranking of archaeological data.

On account of its diverse applications and the intrinsic difficulty of the problem, the QAP has been investigated extensively by the research community. The QAP has been proved to be an NP-complete problem and a variety of exact and heuristic algorithms have been proposed. Exact algorithms for QAP include approaches based on dynamic programming [2], cutting planes [3], and branch and bound [4, 5]. Among these, only branch and bound algorithms are guaranteed to obtain the optimum solution, but they are generally unable to solve problems of size larger than $n=20$. Since many applications of QAP give rise to problems of size far greater than 20, there is a special need for good heuristics for QAP that can solve large-size problems. Known heuristics for QAP can be classified into the following categories: construction methods [6, 7], limited enumeration methods [8, 9], local improvement methods [10], simulated annealing methods [11],

tabu search methods [12, 13] and genetic algorithms [12, 14]. Among these, the tabu search method due to Skorin-Kapov [12] and the (randomized) local improvement method due to Li et al. [10] and Pardalos et al. [15], which they named as Greedy Randomized Adaptive Search Procedure (GRASP), are the two most accurate heuristic algorithms to solve the QAP. Other approaches were used to solve the QAP problem are ant colony algorithm [16, 17], Path Relinking method [18], hybrid GRASP approach with tabu search[19], hybrid Ant Colony approach with Genetic and Hill Climbing [20] and Parallel tabu search [21].

The structure of the reminder of this paper is as follows: Section 2 is an introduction to PSO algorithm. In section 3, we describe the proposed hybrid algorithm based on PSO and Hill Climbing to solve the problem. Section 4 and 5 are dedicated to describe experimental results and paper conclusion respectively.

II. PARTICULAR SWARM OPTIMIZATION

Particle swarm optimization (PSO) is a population based computational technique inspired from the simulation of social behavior of flock of birds. PSO was originally designed and developed by Eberhart and Kennedy [22]. In PSO, a swarm of particle is used to represent the population of candidate solutions. Each particle is a point in the N-dimensional search space. A particle is represented by its current position, and its current velocity. PSO tries to find the optimal solution to the problem by moving the particles and evaluating the fitness of the new position.

III. HYBRID PSO ALGORITHM FOR THE QUADRATIC ASSIGNMENT PROBLEM

Evaluation of each particle in the swarm requires the determination of the permutation of numbers 1...n since the value of function z in QAP problem is a result of the sequence. In this paper, we use a heuristic rule called Smallest Position Value (SPV) to enable the continuous PSO algorithm to be applied to all class of sequencing problems, which are NP-hard in the literature [23]. By using the SPV rule, the permutation can be determined through the position values of the particle so that the fitness value of the particle can then be computed with that permutation. Pseudo code of the PSO algorithm for the QAP is given in Figure 1.

```

Initialize parameters
Initialize population
Find sequence
Evaluate
Do {
    Find the personal best
    Find the global best
    Update velocity
    Update position
    Find sequence
    Evaluate
    Apply hill climbing (optional)
} While (Termination)

```

Fig. 1. The PSO Algorithm with Hill Climbing for the QAP Problem.

The basic elements of PSO algorithm is summarized as follows:

Particle: X_i^k denotes the i^{th} particle in the swarm at iteration k and is represented by n number of dimensions as $X_i^k = [x_{i1}^k, \dots, x_{in}^k]$, where x_{ij}^k is the position value of the i^{th} particle with respect to the j^{th} dimension ($j = 1, \dots, n$).

Population: pop^k is the set of m particles in the swarm at iteration k , i.e., $pop^k = [X_1^k, \dots, X_m^k]$.

Sequence: We introduce a new variable ϕ_i^k , which is a permutation of numbers 1...n implied by the particle X_i^k . It can be described as $\phi_i^k = [\phi_{i1}^k, \dots, \phi_{in}^k]$, where ϕ_{ij}^k is the assigned facility to location j of the particle i at iteration k with respect to the j^{th} dimension.

Particle velocity: V_i^k is the velocity of particle i at iteration k . It can be defined as $V_i^k = [v_{i1}^k, \dots, v_{in}^k]$, where v_{ij}^k is the velocity of particle i at iteration k with respect to the j^{th} dimension.

Inertia weight: w^k is a parameter to control the impact of the previous velocities on the current velocity.

Personal best: PB_i^k represents the best position of the particle i with the best fitness until iteration k . So, the best position associated with the best fitness value of the particle i obtained so far is called the *personal best*. For each particle in the swarm, PB_i^k can be determined and updated at each iteration k . In a minimization problem with the objective function $z(\phi_i^k)$ where ϕ_i^k is the corresponding sequence of particle x_i^k , the personal best PB_i^k of the i^{th} particle is obtained such that $z(\phi_i^k) \leq z(\phi_i^{k-1})$ where ϕ_i^k is the corresponding sequence of personal best PB_i^k and ϕ_i^{k-1} is the corresponding sequence of personal best PB_i^{k-1} . To simplify, we denote the fitness function of the personal best as $z_i^{pb} = z(\phi_i^k)$. For each particle, the personal best is defined as $PB_i^k = [pb_{i1}^k, \dots, pb_{in}^k]$ Where pb_{ij}^k is the position value of the i^{th} personal best with respect to the j^{th} dimension.

Global best: GB^k denotes the best position of the globally best particle achieved so far in the whole swarm. For this reason, the global best can be obtained such that $z(\phi^k) \leq z(\phi_i^k)$ for $i = 1, 2, \dots, n$ where ϕ^k is the corresponding sequence of global best GB^k and ϕ_i^k is the corresponding sequence of particle best PB_i^k . To simplify, we denote the fitness function of the global best as $z^{pb} = z(\phi^k)$. The global best is then defined as $GB^k = [gb_1^k, \dots, gb_n^k]$ where gb_i^k is the position value of the global best with respect to the j^{th} dimension.

Termination criterion: It is a condition that the search process will be terminated. It might be a maximum number of iteration or maximum CPU time to terminate the search.

Solution Representation: One of the most important issue when designing the PSO algorithm lies on its solution representation. In order to construct a direct relationship between the problem domain and the PSO particles for the QAP problem, we present n number of dimensions for n number of locations (j =1,..., n). In other words, each dimension represents a typical location. In addition, the particle $X_i^k = [x_{i1}^k, \dots, x_{in}^k]$ corresponds to the position values for n number of locations in the QAP problem. Each particle has a continuous set of position values. The particle itself does not present a sequence. Instead we use the SPV rule to determine the allocation of facilities to different locations implied by the position values x_{ij}^k of particle X_i^k . Table 1 illustrates the solution representation of particle X_i^k for the PSO algorithm with its velocity and corresponding sequence. According to the proposed SPV rule, the smallest position value is $x_{i5}^k = -1.20$, so the dimension j=5 is assigned to be the first facility $\phi_{i1}^k = 5$ in the sequence ϕ_i^k ; the second smallest position value is $x_{i2}^k = -0.99$, so the dimension j=2 is assigned to be the second facility $\phi_{i2}^k = 2$ in the sequence ϕ_i^k , and so on. In other words, dimensions are sorted according to the SPV rule, i.e., to x_{ij}^k values to construct the sequence ϕ_i^k . This representation is unique in terms of finding new solutions since positions of each particle are updated at each iteration k in the PSO algorithm, thus resulting in different sequences at each iteration k.

TABLE I. SOLUTION REPRESENTATION OF A PARTICLE

j	1	2	3	4	5	6
x_{ij}^k	1.80	-0.99	3.01	-0.72	-1.20	2.15
v_{ij}^k	3.89	2.94	3.08	-0.87	-0.20	3.16
ϕ_{ij}^k	5	2	4	1	6	3

In the above representation, the value of every cell (ϕ_{ij}^k) represents the facility that is assigned to corresponding location. In this example, there are 6 facilities to be placed at 6 locations. For example, the first cell means that facility 5 is placed at location 1; facility 2 is placed at location 2 and so on.

Initial population: A population of particles is constructed randomly for the PSO algorithm of the QAP problem. The continuous values of positions are established randomly. The following formula is used to construct the initial continuous position values of the particle: $x_{ij}^0 = x_{\min} + (x_{\max} - x_{\min}) \times U(0,1)$ where $x_{\min} = 0.0$, $x_{\max} = 4.0$ and U(0,1) a uniform random number between 0 and 1. Initial continuous velocities are generated by

similar formula as follows: $v_{ij}^0 = v_{\min} + (v_{\max} - v_{\min}) \times U(0,1)$ where $v_{\min} = -4.0$, $v_{\max} = 4.0$ and U(0,1) a uniform random number between 0 and 1. Continuous velocity values are restricted to some range, namely $v_{ij}^k = [v_{\min}, v_{\max}] = [-4.0, 4.0]$ where $v_{\min} = -v_{\max}$.

As the formulation of the QAP problem suggests that the objective is to minimize cost Z, the fitness function value for the particle i of the population (or swarm) at the iteration k is,

$$Z_i^k(\phi_i^k) = \sum_{i=1}^n \sum_{j=1}^n f_{ij} d_{\phi_i^k(i)\phi_i^k(j)} + \sum_{i=1}^n c_{i\phi_i^k(i)}$$

Where ϕ_i^k is the corresponding sequence of particle x_i^k . For simplicity, $Z_i^k(\phi_i^k)$ will be denoted Z_i^k . The complete computational flow of the PSO algorithm for the QAP problem can be summarized as follows:

Step 1: Initialization

- Set $k=0$, m =size of swarm.
- Generate m particles randomly as explained before, $\{X_i^0, i = 1, \dots, m\}$ where $X_i^0 = [x_{i1}^0, \dots, x_{in}^0]$.
- Generate initial velocities of particles randomly $\{V_i^0, i = 1, \dots, m\}$ where $V_i^0 = [v_{i1}^0, \dots, v_{in}^0]$.
- Apply the SPV rule to find the sequence $\phi_i^0 = [\phi_{i1}^0, \dots, \phi_{im}^0]$ of particle X_i^0 for $i=1, \dots, m$.
- Evaluate each particle i in the swarm using the objective function Z_i^0 for $i=1, \dots, m$.
- For each particle i in the swarm, set $pB_i^0 = X_i^0$, where $pB_i^0 = [pb_{i1}^0 = x_{i1}^0, \dots, pb_{in}^0 = x_{in}^0]$ along with its best fitness value, $Z_i^{pb} = Z_i^0$ for $i=1, \dots, m$.
- Find the best fitness value $Z_i^0 = \min\{Z_i^0\}$ for $i=1, \dots, m$ with its corresponding position X_i^0 .
- Set global best to $GB^0 = X_i^0$ where $GB^0 = [gb_1 = x_{i1}^0, \dots, gb_n = x_{in}^0]$ with its fitness value $Z^{pb} = Z_i^0$

Step 2: Update iteration counter

- $k = k + 1$

Step3: Update inertia weight

- $w^k = w^{k-1} \times \alpha$ where α is decrement factor.

Step 4: Update velocity

- $v_{ij}^k = w^{k-1} v_{ij}^{k-1} + c_1 r_1 (pb_{ij}^{k-1} - x_{ij}^{k-1}) + c_2 r_2 (gb_j^{k-1} - x_{ij}^{k-1})$

Step 5: Update position

- $x_{ij}^k = x_{ij}^{k-1} + v_{ij}^k$

Step 6: Find Sequence

- Apply the SPV rule to find the sequence $\phi_i^k = [\phi_{i1}^k, \dots, \phi_{im}^k]$ for $i = 1, \dots, m$

Step 7: Update personal best

- Each particle is evaluated by using its sequence to see if personal best will improve. That is, if $\phi_i^k < \phi_i^{pb}$ for $i = 1, \dots, m$, then personal best is updated as $PB_i^k = X_i^k$ and $Z_i^{pb} = Z_i^k$ for $i = 1, \dots, m$.

Step 8: Update global best

- Find the minimum value of personal best $Z_l^k = \min\{Z_i^{pb}\}$ for $i = 1, \dots, m$, $l \in \{i : i = 1, \dots, m\}$
- If $\phi_l^k < \phi^{gb}$, then the global best is updated as $GB^k = X_l^k$ and $Z^{gb} = Z_l^k$

Step 9: Stopping criterion

- If the number of iteration exceeds the maximum number of iteration, or maximum CPU time, then stop, otherwise go to step 2.

Hill climbing for the QAP problem: In the proposed PSO algorithm, Hill Climbing is applied to the sequence directly. However, it violates the SPV rule and needs a repair algorithm. This approach is illustrated in Table 2 and 3, where facility $\phi_{12}^k = 2$ and facility $\phi_{14}^k = 1$ are interchanged.

TABLE II. HILL CLIMBING APPLIED TO SEQUENCE BEFORE REPAIRING

j	1	2	3	4	5	6
x_{ij}^k	1.80	-0.99	3.01	-0.72	-1.20	2.15
Facility, ϕ_{ij}^k	5	2	4	<u>1</u>	6	3
x_{ij}^k	1.80	-0.99	3.01	-0.72	-1.20	2.15
Facility, ϕ_{ij}^k	5	<u>1</u>	4	2	6	3

As seen in Table 2, applying a Hill Climbing to the sequence violates the SPV rule because the sequence itself is a result of the particle's position values. Once a Hill Climbing is completed, particle should be repaired so that the SPV is not violated. This is achieved by changing the position values according to the SPV rule as shown in Table 3.

TABLE III. HILL CLIMBING APPLIED TO SEQUENCE AFTER REPAIRING

J	1	2	3	4	5	6
x_{ij}^k	<u>1.80</u>	<u>-0.99</u>	3.01	-0.72	-1.20	2.15
Facility, ϕ_{ij}^k	5	2	4	<u>1</u>	6	3
x_{ij}^k	<u>-0.99</u>	<u>1.80</u>	3.01	-0.72	-1.20	2.15
Facility, ϕ_{ij}^k	5	<u>1</u>	4	2	6	3

In other words, interchange the position values of the exchanged facilities in terms of their dimensions. Since facility $\phi_{12}^k = 2$ and facility $\phi_{14}^k = 1$ are interchanged, their associated

position values $x_{12}^k = -0.99$ and $x_{14}^k = 1.80$ are interchanged for dimensions $j=2$ and $j=1$ to keep the particle consistent with the SPV rule.

The Pseudo Code of Hill Climbing for the QAP Problem is given in figure 2. In the figure p_{HC} is the probability of Hill climbing.

```

current= $\phi^k$ , sequence of global best  $GB^k$ ;
Iteration=0;
Do
{
  u=rd(1,n);
  min_cost=int.MaxValue;
  for (i = 1; i <= n; i++)
    If (i!=u)
    {
      S=interchange(current, u,i);
      Repaire(s);
      If (Z(s)<min_cost)
      {
        min_cost=Z(s);
        neighbour = S;
      }
    }
  If (min_cost<Z(current)) current= neighbour;
} while (Iteration<pHC*swarmsize);

```

Fig. 2. The Pseudo code of hill climbing for The QAP problem

IV. EXPERIMENTAL RESULTS

In this section the results of the implementation of the new algorithm is described. The hybrid PSO_{HC} algorithm presented for the QAP problem was coded in the C#.Net 2010 programming language.

We used the following parameters for the PSO. Social and cognitive parameters, and uniform random numbers are taken as $c1 = c2 = 2$ and $r1 = r2 = 0.5$ respectively. Initial inertia weight is set to $w0 = 0.9$ and never decreased below 0.40. Finally, the decrement factor α is taken as 0.975. We evaluated the performance of the proposed algorithm using the benchmark set of instances, available on the QAPLIB compiled by Burkard et al. [24]. (<http://mscmga.ms.ic.ac.uk/jeb/orlib/wtinfo.html>).

Experimental results and analysis are divided into two parts. First, in section A, we evaluate the effect of important parameters on the proposed algorithm. In section B, we present the computational results of the hybrid algorithm when applied to some instances of the QAPLIB.

A. Evaluate the effect of parameters on ICA algorithm

It is clear that the performance of the approximated algorithms is affected by parameter tuning. So, at first we do tuning process, to obtain good values for the key parameters.

We use bur26f instance of QAPLIB to test the effect of PSO_{HC} parameters. This instance contains 26 facilities. We increase the number of iterations of the algorithm ranging from 100 to 500. Figure 3 shows the results. The first point is that as the number of iterations increases, the quality of solutions will

improve. There is another noticeable point in which the cost decreased sharply to 400 iterations, then it stabilizes. So, to balance the quality of the results and the running time, we set the maximum number of iterations to be 400 for following experiments.

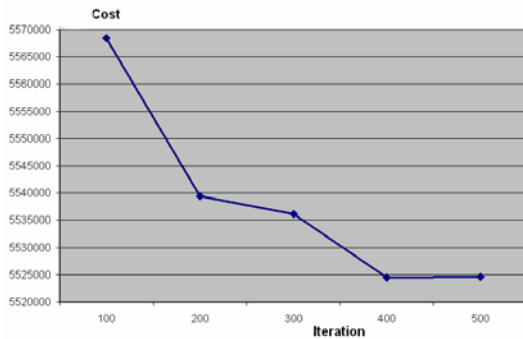


Fig. 3. The impact of increasing the PSO_{HC} 's number of iterations

We choose the size of swarm from the set $\{40, 80, 120, 160$ and $200\}$. The figure 4 shows the results. The figure indicates that increasing the size of swarm makes better solutions. The quality of solutions improves until 120 particles, and then it is stable. So we've used 120 in our experiments.

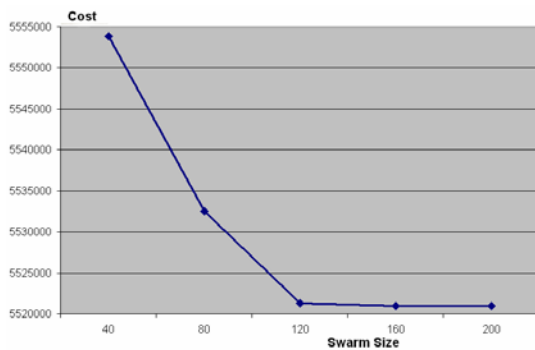


Fig. 4. The impact of increasing the size of swarm

B. Performance of Algorithm

Performance of the pure PSO, denoted as PSO_{pure} , and PSO with the Hill Climbing, denoted as PSO_{HC} , is compared with a simulated annealing method, a kind of genetic algorithm which is called GA_OB and PSO with local Search, denoted as PSO_{VNS} [23]. Table 5 shows the comparisons. The first column contains the QAP instance and second contains the size of problem. The third column shows the Best-Known Solution (BKS). The Solution and gap Column are respectively the solutions produced by each algorithm and the deviation in percentage from BKS.

According to the results, PSO_{HC} is able to find optimal and near optimal solutions. It means that in most cases the results are as well as BKS. In addition to the comparisons indicate that the new approach is better than other methods which have been proposed. So we can use PSO_{HC} approach as an effective method to solve the QAP problem.

V. CONCLUSIONS

The quadratic assignment problem is a very difficult combinatorial optimization problem. In order to obtain satisfactory results in a reasonable time, heuristic algorithms are to be applied. In this paper, we developed a hybrid modified PSO method in which a simple but very efficient Hill Climbing method is embedded.

By comparison the results produced by PSO_{HC} with the results by SA, GA, PSO_{pure} and PSO_{VNS} , we find that PSO_{HC} do better and find the solutions which have low cost. These results are nearly as well as Best-Known Solutions. Experimental results illustrate the effectiveness of proposed approach on the quadratic assignment problem.

REFERENCES

- [1] R.K. Ahuja, J.B. Orlin, and A. Tiwari, "A greedy genetic algorithm for the quadratic assignment problem", *Computers and Operations Research* 27, 917-934. 2000.
- [2] Christodoulos N, Benavent E. An exact algorithm for the quadratic assignment problem. *Oper. Res.* 1989;37:760-8.
- [3] Bazara MS, Sherali MD. Benders' partitioning scheme applied to a new formulation of the quadratic assignment problem. *Naval Res. Logist. Quart.* 1980;27:29-41.
- [4] Lawler EL. The quadratic assignment problem. *Manag. Sci.* 1963;9:586-99.
- [5] Pardalos PM, Crouse J. A parallel algorithm for the quadratic assignment problem. *Proceedings of the Supercomputing 1989 Conference*, ACM Press. New York, 1989, pp. 351-60.
- [6] Armour GC, Bula ES. Heuristic algorithm and simulation approach to relative location of facilities. *Manag. Sci.* 1963;9:294-309.
- [7] Bula ES, Armour GC, Vollmann TE. Allocating facilities with CRAFT. *Harvard Business Rev.* 1962;42:136-58.
- [8] West DH. Algorithm 608: approximate solution of the quadratic assignment problem. *ACM Trans. Math. Software* 1983;9:461-6.
- [9] Burkard RE, Bonniger T. A heuristic for quadratic boolean programs with applications to quadratic assignment problems. *European J. Oper. Res.* 1983;13:374-86.
- [10] Li T, Pardalos PM, Resende MGC. A greedy randomized adaptive search procedure for the quadratic assignment problem. In: Pardalos PM, Wolkowicz H (Eds.), *Quadratic Assignment and Related Problems*. DIMACS Series in Discrete Mathematics and Theoretical Computer Science. Providence, RI: American Mathematical Society, 1994, pp. 237-261.
- [11] Wilhelm MR, Ward TL. Solving quadratic assignment problems by simulated annealing. *IEEE Trans.* 1987;19:107-19.
- [12] Skorin-Kapov J. Tabu search applied to the quadratic assignment problem. *ORSA J. Comput.* 1990;2:33-45. R.K. Ahuja et al. *Computers & Operations Research* 27 (2000) 917-934.
- [13] Taillard E. Robust tabu search for the quadratic assignment problem. *Parallel Comput.* 1991;17:443-55.
- [14] Fleurent C, Ferland JA. Genetic hybrids for the quadratic assignment problem. *DIMACS Series in Discrete Mathematics and Theoretical Computer Science*, vol. 16, Providence, RI: American Mathematical Society, 1994, pp. 173-87.
- [15] Resende MGC, Pardalos PM, Li Y. Fortran subroutines for approximate solution of dense quadratic assignment problems using GRASP. *ACM Trans. Math. Software* 1994;22:104-18.
- [16] Stutzle, T., Dorigo, M., 1999. ACO algorithms for the quadratic assignment problem. In: Come, D., Dorigo, M., Glover, F. (Eds.), *New Ideas for Optimization*. McGraw-Hill, pp. 33-50.

- [17] Gambardella, L., Taillard, E., Dorigo, M., 1997. Ant Colonies for the QAP, Tech. Report IDSIA-4-97, IDSIA, Lugano, Switzerland.
- [18] James, T., Rego, C., Glover, F., 2005. Sequential and parallel pathrelinking algorithms for the quadratic assignment problem. *IEEE Intelligent Systems* 20 (4), 58–65.
- [19] Oliveira, C.A., Pardalos, P.M., Resende, M.G.C., 2004. GRASP with pathrelinking for the quadratic assignment problem, Efficient and Experimental Algorithms. In: Ribeiro, C.C., Martins, S.L. (Eds.), . In: *Lecture Notes in Computer Science*, vol. 3059. Springer-Verlag, pp. 356–368.
- [20] Tseng, L., Rego, C., Glover, F., 2009. A cooperative parallel tabu search algorithm for the quadratic assignment problem. *European Journal of Operational Research* 195, 810–826.
- [21] James, T., Liang, S., 2005. A hybrid metaheuristic for the quadratic assignment problem. *Computational Optimization and Applications* 34, 85–113.
- [22] J. Kennedy and R. C. Eberhart, “Particle Swarm Optimization,” in *IEEE International Conference on Neural Networks*, Piscataway, NJ, 1995, pp. 1942-1948.
- [23] M. Fatih Tasgetiren, Mehmet Sevkli, Yun-Chia Liang, Gunes Gencyilmaz, “Particle Swarm Optimization Algorithm for the Single Machine Total Weighted Tardiness Problem”, In: the Proceeding of the World Congress on Evolutionary Computation (CEC2004), 2004, p.1412-1419.
- [24] Burkard RE, Karisch SE, Rendl F. QAPLIB - A quadratic assignment program library. *J. Global Optim.* 1997;10:391-403.

TABLE IV. PERFORMANCE COMPARISON ACCORDING TO PROBLEM SIZE

Problem	N	BKS	PSOPure		PSOVNS		PSOHC		SA		OB-GA	
			Solution	Gap(%)	Solution	Gap(%)	Solution	Gap(%)	Solution	Gap(%)	Solution	Gap(%)
tai10a	10	135028	135028	0.000	135028	0.000	135028	0.000	135028	0.000	135028	0.000
tai20a	20	703482	704190	0.101	703482	0.000	703482	0.000	703482	0.000	703482	0.000
tai30a	30	1818146	1868871	2.790	1866122	2.639	1848571	1.673	1849696	1.735	1863722	2.506
tai40a	40	3139370	3233491	2.998	3225382	2.740	3211400	2.294	3212692	2.335	3215382	2.42125012
tai50a	50	4941419	5088574	2.978	5043551	2.067	5040012	1.995	5041058	2.016	5067904	2.559
tai60a	60	7208572	7438398	3.188	7431459	3.092	7325133	1.617	7381442	2.398	7422318	2.965
tai80a	80	13557864	13831255	2.016	13812280	1.877	13800423	1.789	13852100	2.170	13814562	1.893
tai100a	100	21125314	21603991	2.266	21602881	2.267	21483465	1.695	21499478	1.771	21501554	1.780
tai150b	150	498896643	508956543	2.016	506448890	1.514	502349984	0.692	502651565	0.572	506448890	1.513
tai256c	256	44759294	44819486	0.134	44809949	0.113	44799323	0.089	44814014	0.122	45023121	0.589
sco100a	100	152002	155462	2.276	153949	1.281	152922	0.605	154210	1.452	153090	0.715
sco100b	100	153890	156007	1.376	155971	1.352	154123	0.151	154262	0.241	155030	0.740
sco100c	100	147862	150978	2.107	149366	1.017	149742	1.271	149542	1.136	149948	1.410
sco100d	100	149576	152346	1.852	151746	1.451	150431	0.572	151746	1.450	150828	0.837
sco100e	100	149150	152349	2.145	150999	1.240	151426	1.526	150426	0.855	150598	0.970
sco100f	100	149036	152038	2.014	150871	1.231	150111	0.721	150738	1.142	150402	0.916

The Usage of Malay Technological Terminologies in Malaysian Youth Institutions for Skills: Interests and Challenges

Adenan Ayob
Sultan Idris Education University
35900 Tanjung Malim
Perak, MALAYSIA
adenan@fbk.upsi.edu.my

Abstract - A very common problem of terminology work is the importance and indeed the very nature of terminology that poorly understood. Thus many people simply have no idea at all of what it is, while others searching for an explanation of some sort, end up associating it with different meaning. Related professions in the communications field, such as translation and technical will often be aware of the word without having precise knowledge of what it entails. This paper highlights the interests and challenges of the usage of Malay technological terminologies in Malaysian youth institutions for skills. Other than discussing the meaning of terminology and technology and their concept, this paper recommends the proposed methodology for the development of technology terminologies that based on the theory of Post-Modernism and Constructivism. That theory has made coverage of scientific research for International Energy Agency. The main aims under those discussions were to discover and to ensure that the terminologies to be developed to function automatically and to accelerate the collection process, the creation and dissemination in information and communication technology.

Keywords: Malay technological terminology, skills, interests, challenges, ICT

I. INTRODUCTION

Before pioneering the interests and challenges, the main problem is related to the area of using foreign terminologies, especially the terminologies that borrowed and modified from other languages. Notably, the initiative to introduce new terminologies in Malay language technology has been implemented for a long time since the establishment of Malaysia National University in 1970. The implementation of technological programs in Malaysia National

University covers many important aspects of terminology.

The fact that is desired, the Malaysian youth institutions for skills can refer to various resources that based on the technological terminologies from dictionary and glossary based on information and communication technology as published by the Malaysian government. The portal that organized by the government too is a popular reference at this time to ensure that the terminologies used are accurate. However, the actual reference material produced specifically for Malaysian youth institutions for skills should be issued regarding certain technological courses.

With a variety of platforms that can be accessed by each student at the institute, either through internet or printed materials, the issue of an inappropriate use of terminologies will not arise again. Steps taken through creation of technological terminologies were to cover a lot of problems on the widespread of foreign terminologies. This gives the impression that the usage of technological terminologies is more important.

There's no denying that the use of more precise technological terminologies in those institutes may be awkward for some people. However, the responsibility of accurate terminology is the role of all parties. The management and students of those institutions should be more sensitive to the development of terminology. In fact, the reference could be easily accessed, especially with the versatility of information and communication technology, computers and smart phones available.

Many benefits can be achieved through technology equipment. Area of internet access has also been established in Malaysia. Thus, when faced with a situation to choose the right terminologies, students can easily reach relevant portal for reference. Other alternative might be through dictionary and glossary references.

Dictionary and glossary, reference should be made in the selection of an appropriate and accurate terminology.

II. TECHNOLOGY

Technology, according to Hervey and Higgins (1992) is closely linked with science and engineering. In other words, the technology consists of two dimensions, namely science and engineering that are interconnected with each other. Science refers to understanding the real world in the human environment. This means that technology is a basic feature in the dimensions of materials and power and the interaction of science.

In fact, according to Brislin and Richard (1976), science is a medium that ultimately create a culture and an expression of physics. According Brislin and Richard too, the engineering is a matter connected with knowledge about materials and power that applied in the areas of planning, including technical matters. In another sense, technology accounts for techniques and tools for maintaining a work that based on the results of science.

III. TECHNOLOGICAL TERMINOLOGIES CONCEPTS

Technology by Catford (1965) conceptually have three principles, namely (i) technology; human artifacts, including the hardware and the system of large-scale complex technology, (ii) the nature of technological creation and discovery, development and dissemination to the public widely, and (iii) technology, which began from a very specific technique and practically involving scientific technology systems. Therefore, technology is defined as the study of the relationship between human and the world, which manifests itself in view of technology, research on the phenomenon of the overall technology, placement of technology in community development retrospectively and prospectively in accordance with those dimensions.

According to technology research purposes by Brislin and Richard (1976), it is focused on the technical sciences or engineering, technical products, activities and knowledge as a cultural phenomenon. Brislin and Richard added that technology involves technical development and knowledge.

Similar views by Nida, Eugene A (1964), it is also related to technological progress and its relationship with the philosophy of technology. Nida, Eugene A divides the philosophy of technology in two stages, namely the cognitive and instrumental. Each stage will be followed by rapid technology change and vary.

IV. THE IMPORTANCE OF USING MALAY TECHNOLOGICAL TERMINOLOGIES IN MALAYSIAN YOUTH INSTITUTIONS FOR SKILLS

The technological terminology in Malaysian youth institutions for skill was centered on the importance of an underlying scientific basis and fundamental science. These interests include applied science and applied research. Actual sustainability terminologies should be seen as two bands that are complementary and generate the ideas that based on tools and materials.

In cognitive aspect, knowledge that based on technology is designed efficiently to resolve practical problems (Arrowsmith & R. Shattuck, 1961). Technology changes the capacity to apply scientific aspects in the development of technological knowledge.

In instrumental level, technology is a set of artifacts that are designed and produced intensive to perform the functions of mechanical and electronic (Arrowsmith & R. Shattuck, 1961). Changes in the instrumental-based technology is consistent and in line with the aspirations of the development of technology. The underlying technology of the entity is not a system of knowledge, but rather a complex system designed from the intentional operator.

Technology is the medium of terminology resources that easily and quickly manage to acquire knowledge. Awareness of the importance of terminology has prompted adoption of skills needed by students to compete in the new millennium.

Literacy in terminologies of learning technologies can purpose efficiently and effectively. So, rapid progress in technology allows a student to collect, transmit, distribute, manage, process or store various types of information quickly and easily (Bell, 1911). Technology is considered a new facility in the smart education system.

The usage of the terminologies in youth institution for skill is the first step towards creating a technological society in line with the progress of the country. The terminologies that controlled technology allowed students to easily master the knowledge and skills in order to use the facility for the challenges in industrial world.

V. THE CHALLENGES OF USING MALAY TECHNOLOGICAL TERMINOLOGIES IN MALAYSIA YOUTH INSTITUTION FOR SKILLS

Among those challenges, the terminologies that exist in communications felt awkward when attempted to be preserved. This was probably due to an inappropriate interpretation of such terminologies. Such complexity also gives effect to the interpretation. Interpretation involves a source language in the recipient. In this process, it should be identified by the expression equation method based on the meaning of the source language and style or the way of common language. The aim is to improve the ability of consumers to make timely and accurate interpretation based on the formation of a creative and innovative thinking.

The above statement reflects the existence of problems in any absolute balance. An axis of problems often emerged from the resolution of terminologies due to constraints of remuneration, strings, and the coefficient in plural noun. Therefore, the importance of interpretation for such terminology has not been formulated as a whole. In fact, to show any strong interpretation of terminologies should be consistent with a principle or theory.

This means that there is no expression on the interpretation of terminologies to suit a model equation of language sources. In other sense, all kinds of interpretations of terminologies may be an addition and distortion of information.

Clearly, in order to interpret accurately and neatly on the barrel of technical terminologies, a complex problem will always exist. Thus, this study will focus on a suitable method that is characterized by the terminology development theory based, including the interpretation under practical facilities. The terms may not be able to carry the intended meaning. Therefore, the interpretation of terminologies should be attributed to the expression of needs-based communication process.

The study to be conducted was focused on the attitude that based on the dimensions of (i) perceptual; knowledge and confidence, (ii) affective; facilitate and (iii) behavior. The practical dimensions are involved (i) strategies, (ii) the level and (iii) interest, while a poll among experts will focus on the content and appropriateness of the terminology which is based on the curriculum and syllabus, and the importance of practice and interpretation procedures, including practicalities study that based upon a degree course.

VI. PROPOSED METHODOLOGY

Recommendations proposed methodology for the development of technology terminologies is based on the theory of Post-Modernism and Constructivism. That theory has made coverage of scientific research for International Energy Agency which is adapted to the theory of Post-Modernism and Constructivism. In the countries concerned with several components of the source terminologies development in technology, students tend toward the formation of attitudes and practices of the sustainability of a technical course. Therefore, it can be justified that the youth in Malaysia should realize the vision for the interpretation of terminologies might think the importance of technology in building strong technical thinking.

Theory is implicit in the concept of sustainability accuracy of an interpretation towards terminology. According to The Partnership for a Combination of 21st Century Skills or P21, the accuracy of an interpretation is a concept of sustainability in the appearance of a technical knowledge that to be mastered. This includes the ability for a student to master the strategies and skills related to creative and innovative thinking, as well as other skills across towards establishing autonomy in thinking. Intended interpretation of the concept of accuracy is not limited to the control of terminologies, but also the feasibility and application of technical skills to meet all demands for the development of an industry. The command interpretation for the 21st century youth should give the ability to mutually combine local knowledge with global knowledge and make them flexible and efficient to adapt the form of knowledge and also sharing knowledge that is constantly changing and do not marginalized or left behind.

VII. CONCLUSION

Under the technological terminology, the work of lexicography should be seen as an important role in generating balance language and technology development. To compete with other foreign terminologies, a development of more innovative methods deemed appropriate. This effort is at least in the context of terminologies collection that to be implemented. The proposed methodology involves the collection of texts relating to specific areas of the experimental identification of terminologies under the supervision of local experts and related to the perform analysis of such courses.

Therefore, the usage of the terminologies that regards upon research at these institutions should be seen as an early step in the construction of technological system. The aim is to ensure that the terminologies to be developed to function automatically and to accelerate the collection process, the creation and dissemination.

ACKNOWLEDGEMENT

This research is my original work. The research contributes to the acquisition of terminologies and technological concepts.

REFERENCES

- [1] Arrowsmith, W., and R. Shattuck, *The craft and context of translation*. The University of Texas Press, 1961.
- [2] Bell, R.T., *Translation and translating: Theory and practice*. London: Longman, 1911.
- [3] Brislin, Richard W. (ed.), *Translation: Application and research*. New York: Gardner Press, 1976.
- [4] Catford, J.C., *A linguistics theory of translation*. London: Oxford University Press, 1965.
- [5] Hervey, S. Higgins, I., *Thinking translation*. London: Routledge, 1992.
- [6] Nida, Eugene A., *Toward a science of translation*. Leiden: E.J. Brill, 1964.

Computing Infrastructure and Services Deployment for Research Community of Moldova

P. Bogatencov, G. Secrieru, N. Iliuha
RENAM Association and
Institute of Mathematics and Computer Science of ASM
5, Academiei, Chisinau, MD-2028, Moldova

Abstract— Scientific Computing Infrastructure, related technologies and services have begun developing for R&D communities of Moldova due to the support of a series of international and national projects. These projects are focused on forming the national computational facilities as a part of national, regional and European eInfrastructures. Their aim is to provide access to modern computational resources to wide range of researchers, attract new research communities in Moldova and promote joint research activities at national, regional and European levels. The realization of eInfrastructure development projects will allow national research communities to get access to the computing resources of leadership-class capability and will contribute to the competitiveness of national research teams.

I. INTRODUCTION

The transition of the traditional science to e-Science is fueled by the ever-increasing need for processing exceedingly large amounts of data and exponentially increasing computational requirements. In order to realistically describe and solve real-world problems, numerical simulations are becoming more detailed, experimental sciences use more complicated instruments to make precise measurements. Now the shift from the individuals-based science work towards collaborative research model starts to dominate. In this context the role of Scientific Computing (HPC, Grids, Cloud computing) in the modern scientific research is crucially increasing. It considerably determines the level of development of the scientific knowledge based society. Mathematical modeling forms a solid theoretical and applied basis in describing, simulating and studying the complex problems. The international cooperation in the field of Scientific Computing represents an important factor for developing the area of scientific research and perspectives of the European future for research community of Moldova.

In the last years Moldova as a part of South-East Europe actively participated in a number of targeted initiatives funded by the European Commission, focused on the creation of new user communities, and enabling collaborative research across some fields in South-East Europe. Although the necessary initial contributions in the region were done, the computational facilities available now are in general less developed than in Western Europe [1]. Advancing the Information Society in such countries as Moldova, strengthening the local eInfrastructures, activating new user communities and enabling collaborative research across a number of fields, would

strongly contribute to closing the existing technological and scientific gap, and thus bridging the digital divide, stimulating and consequently alleviating the brain drain in the region of South-East Europe.

II. APPROACHES OF COMPUTATIONAL INFRASTRUCTURE AND TECHNOLOGIES DEVELOPMENT

Modern Scientific Computing infrastructure is based on using parallel architectures specialized on running complex applications and integrates the following components:

- HPC - Clusters' systems;
- HPC - Supercomputers;
- Distributed computing infrastructure – Grids;
- Scientific Clouds... (for perspective);
- Parallel algorithms design and programming
- Instruments for complex applications development.

Historically in Moldova, the first scientific computing resources had begun developing from the initial deployment in 2006 of the first Grid cluster that was integrated in the regional South-Europe Grid infrastructure.

These specific and new for Moldova activities were supported by a range of the SEE-GRID projects [2]. These projects have allowed to establish strong human network in the area of scientific computing and have set up a powerful regional Grid infrastructure. This is very much in line with the European vision of moving towards a long-term sustainable European Grid Initiative through strong support of National Grid Initiatives (NGI), and, in this aspect, SEE-GRID be to some extent leading the way by its successful establishment of NGIs in the region, including Moldova. One of main objectives of the SEE-GRID projects was to penetrate and engage regional and national user communities via multi-disciplinary grids, involving a range of research and academic institutes and scientific communities in all SEE countries, with emphasis on the deployment and support of a range of Grid applications. RENAM Association, representing national scientific-educational network of Moldova (NREN), started to build MD-Grid NGI: National Grid-Initiative of Moldova Consortium, that was created within the framework of the SEE-GRID-2 project supported by the European Commission [2]. MD-Grid - National Grid Initiative of Moldova was inaugurated on May 14, 2007 after receiving approval letters from Ministry of Information Development of Moldova and the Academy of

Sciences of Moldova. The MD-Grid NGI Consortium is governed by RENAM as its Coordinating body and joins seven research, education and industry institutions that expressed their intent to participate in the processes of building National Grid Infrastructure and using its resources. The major steps in the establishment of MD-Grid NGI included:

- the development and approval of MD-Grid NGI Foundation document (Consortium Agreement),
- the development and approval of MD-Grid NGI Policy document,
- the selection of and negotiation with potential MD-Grid NGI members,
- signing the agreements about intentions and MoU with MD-Grid NGI potential members,
- planning grid infrastructure enlargement, creation and setting up at least three Grid-sites in research and higher educational institutions of Moldova at the beginning stage.

The elaborated MD-Grid NGI Policy document stipulates basic principles of NGI Consortium behavior and requirements for NGI supported services:

- NGI membership rules. NGI member organizations should belong to four main categories: research, academia, industry and government.
- Accessible Use Policies and Services Level Agreements for resources and services.
- National Grid Certification (Registration) Authority policy and behavior.
- Core services structure and functions.

The first accumulated experience was successful from the point of view of forming professional team of specialists in the area of distributed computing and examination of potential users' communities needs in computational resources that pave the way for creation of prepared national users community in future. The main activity directions of the created in Moldova National Grid Initiative are summarized as followed:

- MD-Grid NGI participates in strategic European Programs for the development of transnational grids and in initiatives for the completion of SEE eInfrastructures. The operation of the MD-Grid NGI implements the general EU policy on the development of national initiatives for the coordination of actions related to eInfrastructures and especially to scientific computing infrastructures.
- The integration of Grid actions (infrastructures, middleware and applications) with the broadband network into a standard e-Infrastructures system. The optimization of exploitation of advanced network resources and services, which can serve the new e-Science generation and will attract the greater users' community of the information society to the mass adoption of advanced services provided by Grid architectures.
- The permanent development and administration of the National Grid infrastructure.

- The organization access for national users' communities to the regional and European computational resources (HPC, Grid, scientific clouds, etc.).
- The educational and training events organization; the technological support of national users' communities.

NGI is engaged in the development of Grid infrastructure in Moldova. At present Grid infrastructure unites three sites and has well determined perspectives for its further enlargement. Another principal task that is in focus of NGI is monitoring the research and educational community needs and attracting new research teams that have requirements in complex applications development and in access to special computing resources [3].

The development of national and regional scientific Grid infrastructures is coordinating by pan-European initiatives like EGI-InSPIRE project that is focused on supporting transition process from a project-based system (the EGEE series) to a sustainable pan-European e-Infrastructure. EGI-InSPIRE activities cover grids of high-performance computing (HPC) and high-throughput computing (HTC) resources. The project integrates new Distributed Computing Infrastructures (DCIs) such as clouds, supercomputing networks and desktop grids, to benefit the user communities within the whole European Research Area.

III. ACCESS TO THE REGIONAL HPC RESOURCES AND STRATEGY OF THEIR DEVELOPMENT

Contacts with members of the national research community and analysis of their needs had shown that there are special needs in computational resources that are not covered by Grid infrastructure. That is why MD-Grid NGI argued the necessity, initiated and actively participated in new regional scientific computing activities like HPC and cloud computing.

In the field of High Performance Computing, the European Commission supports a series of initiatives to provide access to HPC facilities to leading European researchers. The SEE region is still lagging behind the European developments in the HPC area. Only few HPC installations are available – one large supercomputing top500 installation in Bulgaria and some smaller HPC ones in a couple of other countries, and these are not open to cross-border research. The user communities that use HPC are limited. Similarly, the less-resources countries have not established any mechanism for interfacing to European HPC infrastructures like PRACE, DEISA, or any other related initiatives. Thus, the regional eInfrastructure must be expanded to address these specific needs of scientific communities in the region. Furthermore, the region has a strong need to acquire and maintain expertise in the provision and utilization of HPC facilities both at the system as well as at the software level.

To cover the permanently rising needs of researchers in SEE region the regional eInfrastructure development project “High-Performance Computing Infrastructure for South East Europe’s Research Communities (HP-SEE)” was elaborated and proposed for funding. The HP-SEE project (<http://www.hp-see.eu/>) started in September 2010 and brings together 14 partners from the SEE region, while more than 10 institutions have been involved in the project as third parties.

The project has begun with only few HPC installations available, being not open to cross-border research. The aim of the South-East Europe HPC initiative is the equal participation of all countries of the region in European eInfrastructure development trends.

HP-SEE focuses on a number of strategic actions [4]:

First, it will link the existing and upcoming HPC facilities in the region in a common infrastructure, and will provide operational solutions for it.

Second, it will open this HPC infrastructure to a wide range of new user communities, including those of less-resourced countries, fostering the collaboration and providing advanced capabilities to researchers, with an emphasis on strategic groups in computational physics, chemistry and life sciences.

Finally, it will ensure the establishment of national HPC initiatives. HP-SEE will aim to attract local political & financial support for long-term sustainable eInfrastructure.

RENAM Association (National Research and Educational Network of Moldova) and the Institute of Mathematics and Computer Science of the Academy of Sciences of Moldova (IMI ASM) are involved in the project from Moldova. RENAM efforts are emphasized on the promotion of national communities to the use of the regional infrastructure for high performance computing, training activities, applications porting and operational support. The main task of IMI ASM is the development of HPC applications and the deployment of them in the regional HPC infrastructure.

In the project there are two categories of partners, which form the project consortium – partners that have their own HPC resources, so called “resource providers” and partners-beneficiaries. Beneficiary countries like Moldova receive preferences from gaining access to resources available in the other project partners in the region - “resource providers”.

The regional HPC infrastructure integrates the most powerful HPC clusters and supercomputers provided by the main infrastructure partners from the six countries, participating in the project: Greece, Bulgaria, Romania, Hungary, Serbia, and FYROM (Macedonia).

The structure of the regional HPC infrastructure is heterogeneous, comprising supercomputers, Intel/AMD CPU and GPU clusters. HPC resources available for users’ community from Moldova include Blue Gene/P supercomputer deployed at Executive Agency “Electronic Communications Networks and Information Systems” in the Bulgarian Supercomputing Centre (BGSC), consisting of two racks, 2048 PowerPC 450 based compute nodes, 8192 processor cores and a total of 4 TB random access memory. There is also possibility to run jobs in HTC mode (High Throughput Computing). Another resource is the HPCG cluster located in IICT of the Bulgarian Academy of Sciences. It has 576 computing cores organized in a blade system. The storage and management nodes have 128 cores. There is an agreement with the West University of Timisoara (Romania) concerning access of Moldavian researchers to the Blue Gene/P supercomputer, the installation of which has been finished recently.

The main resources of the regional HPC infrastructure support parallel programming paradigms like MPI and OpenMP. Most of them also offer possibility to run jobs in HTC mode.

TABLE 1.
HP-SEE INFRASTRUCTURE CURRENT STATUS AND PLAN OF DEVELOPMENT

Country	TFlops			
	2010	2011	2012	2013
Greece	0	0	40	80
Bulgaria	25	31+8GPU	31+20GPU	40+20GPU
Romania	10	26+4GPU	30+20GPU	30+20GPU
Hungary	1	48	48+12GPU	48+12GPU
Serbia	6	6	20	20
Overall	42	111+12GPU	169+52GPU	218+52GPU

Partners participating in the HP-SEE project have established HPC training infrastructure as the prerequisite for the successful organization of various training events. In addition, this infrastructure is usually utilized for applications elaboration, testing and debugging. This infrastructure represents HPC resources provided by partners, which can be used during the practical trainings sessions. In the Table 2 current training infrastructure is shown, but this list will be extended, as additional HPC equipment is being procured and made accessible for HP-SEE partners. Besides the hardware listed in Table 2, for the purpose of organization of larger training events, HP-SEE HPC resource centers can reserve part of its main infrastructure for the practical training sessions organization.

During the life time of the project the enlargement of the existing training infrastructure was produced in Moldova. IMI-RENAM Grid cluster (8 servers) was fully transferred to the virtualization platform Citrix XenServer. All servers now are running the latest version of virtualization software. The resources of this cluster are permanently developed to fit the needs of domestic applications development, testing and debugging. For this purpose additional reconfiguration of Windows Compute Cluster (WCC) 2003 software was made: WCC was entirely transferred on the virtualization platform; current WCC 2003 Cluster configuration is - 25 cores on 7 servers, max 4 cores per one Virtual Machine.

All applications that are developing in the project grouped within the three Virtual Research Communities (VRC): Computational Physics (CP), Computational Chemistry and Life Sciences Virtual Research Community.

Among CP applications adapting on the regional HPC infrastructure there is AMR_PAR application (parallel algorithm and program for solving continuum mechanics equations using Adaptive Mesh Refinement), developed in the Institute of Mathematics and Computer Science of the Academy of Sciences of Moldova. The AMR_PAR application

is considering a continuum mechanics problem, and namely the problem of modeling the explosion of a supernova type II and,

for this example, created the algorithm and parallel program using the AMR method [4].

TABLE 2.
AVAILABLE RESOURCES FOR TRAINING, APPLICATIONS ELABORATION, TESTING AND DEBUGGING

Country	Partner	Number of Nodes	Number of Cores	CPU Architecture	Interconnection	Batch System
BG	IICT	4	1920	GPU/NVIDIA	2xGigabit Ethernet	Torque
RO	UVT	50	400	x86_64	QDR 4xInfiniband	SLURM
RO	UPB	48	544	X86_64/Cell	4xGigabit Ethernet QDR 4xInfiniband	Sun Grid Engine
RO	ISS	4	2100	GPU/Fermi NVIDIA	2xGigabit Ethernet	Rocks Clusters
RS	IPB	2	16	x86_64 2.0GHz	Gigabit Ethernet Infiniband	Torque
RS	IPB	2	16	POWER6 4.0GHz	Gigabit Ethernet Infiniband	Torque
RS	IPB	2	16	PowerXCell 8i	Gigabit Ethernet Infiniband	Torque
RS	IPB	1	16	Nehalem	Gigabit Ethernet Infiniband	Torque
BA	UOBL ETF	2	16	x86_64	Gigabit Ethernet	Torque
MD	RENAM	1-6	8-20	x86_64	2xGigabit Ethernet	CCS2003
AM	IIAP NAS RA	6	48	x86_64	Gigabit Ethernet	Torque
AM	IIAP NAS RA	24	48	x86_64	Gigabit Ethernet	Torque
AM	IIAP NAS RA	1	240	Tesla 1060	GPU	

This method can be applied to any other nowadays problem of continuum mechanics - to calculate the aerodynamics of aircraft, the calculations of the air flow of cars, a large number of other problems of mathematical modeling – the calculation of the flow of blood through the vessels, the calculations of the heart valves, etc. In all these cases, at the beginning of the problem we define a way to highlight areas in which we need to construct the grid, then the program builds a sequence of grids and makes a decision on them. The social impact depends on the problem to be solved, the use of AMR_PAR being of interest for heavy industry (e.g. car body design and development, aircraft aerodynamics), or for healthcare industry.

For every application at least two specific computational recourses (home clusters) available in the regional HPC infrastructure are assigned. The home clusters for AMR_PAR application are the SGI UltraViolet 1000 supercomputer (1152 CPU, 6057 GByte of memory) at the National Information Infrastructure Development Institute located in Pecs, Hungary and HPCG cluster, located at the Institute of Information and Communication Technologies of the Bulgarian Academy of Sciences (576 computing cores; the storage and management nodes have 128 cores).

IV. PERSPECTIVES OF SCIENTIFIC CLOUD COMPUTING INFRASTRUCTURE DEPLOYMENT

The development of scientific clouds is a rather new, but perspective direction of computational technologies

development. For SEE region the needs in cloud computing technologies deployment were analyzed during execution of SEERA-EI project (“South East European Research Area for eInfrastructures”) funded by EC ERA-NET Programme. The carried out analysis had shown strong interest of the regional and especially Moldovan research communities in scientific clouds technologies deployment. In many countries, and Moldova is one of them, clouds are supported by governmental strategies as a perspective technological approach for providing wide range of e-government services. In 2010 the Government of Moldova launched the national initiative known as “M-Cloud Initiative: Providing IT Services for Society.” The Initiative consists of and covers many compartments, directions and projects. Key points of the M-Cloud Initiative are:

- All citizens (including research and educational community) deserve high quality e-services operated by a modern, reliable and cost efficient platform.
- e-Services provided by M-Cloud respond for the satisfaction of all society needs and Government demands.
- M-Cloud architecture is based on the leading edge Cloud Computing technology and service-oriented architecture.
- Modernized ICT regulatory framework may speed up Moldova’s integration into European Union.
- The Initiative supports the implementation of e-Transformation Agenda that is the national priority.

M-Cloud presumes the fulfillment of optimized system designing approaches and technological solutions:

- Based on coherent high-level system architecture;
- Offering comprehensive set of e-services and governmental notifications;
- Focusing on interoperability, optimal and reliable technological solutions;
- Based on utilization of shared technological platforms.

All electronic public services will be operated by M-Cloud – common platform build on cloud computing technological basis. M-Cloud will integrate and combine a number of different types of specialized private clouds. As an example of the specialized private (or more correct “hybrid”) cloud we consider the perspective deployment of scientific cloud infrastructure. M-Cloud centralized resources will be provided by Governmental private cloud (G-Cloud) by delivering many of common services at IaaS, PaaS and SaaS levels. G-Cloud will host and deliver the most priority and widely demanded e-services.

For finding optimal solutions for the deployment of scientific clouds in SEE region, investigations within SEERA-EI project were carried out. On the base of these work it was proposed to launch regional Pilot Call for projects in the area of scientific cloud computing. The Call was recently announced (<http://www.seera-ei-pjc.asm.md/>) and one of its priority topics is the feasibility of the study of approaches for scientific clouds integration to the announced e-government cloud infrastructures in the region.

The study of available and developing in Europe clouds platforms for non-commercial use had shown that there are many offers developing due to the support of various projects. We examined results of some project that had aim to elaborate technological solutions acceptable for open cloud infrastructure deployment. We obtained some useful and interesting outcomes for practical utilization after the analysis of the following projects related to scientific cloud computing:

- Enabling Clouds for e-Science (ECEE) Initiative, supported by NRENs and other partners in Europe, based on close collaboration on Grid computing, to extend into cloud computing. ECEE has the aim to leverage national cloud infrastructures for Europe – there is a series of wide projects, like EGEE (Enabling Grids for e-Science). The Initiative comprises VENUS-C project co-funded by European Commission, as one of six European Distributed Computing Infrastructures (DCIs). VENUS-C combines experiences in Grid infrastructures and Cloud computing to capitalize on EU investments. VENUS-C brings together 14 European partners and supports the following basic research disciplines: biomedicine - integrating widely used tools for bioinformatics, system biology and drug discovery into the VENUS-C infrastructure; data for Science - integrating computing resources through VENUS-C on data repositories.
- StratusLab - FP7 funded project that unites organizations from France, Greece, Switzerland, Spain and Ireland. The project enhances Grid infrastructure with virtualization and cloud technologies. The focus is on developing a complete,

open-source cloud distribution that allows grid and non-grid resource centers to offer and to exploit an “Infrastructure as a Service” cloud.

- GRNET eScience cloud project that has the aim to offer virtualization and storage services for the Greek scientific community. The project strategy is gradual offering of services, starting with shared storage, moving to VM on demand, and then SaaS. The project policy background is existing MoU in place for Grid computing that now is expanding for HPC as well.
- OpenNebula.org initiative that is an open-source project aimed at building the industry standard and open source Cloud computing tools to manage the complexity and heterogeneity of distributed data centers infrastructures.

The main goal of the analysis of existing initiatives in the area of Scientific Clouds is the determination of optimal technological solutions for: finding standard platform to facilitate the interoperability and to support the interactions of Governmental Clouds like M-Cloud and possible open source Scientific cloud solutions; the integration of existing virtualization layers with the on-demand delivery model typical for commercial clouds; the virtualization and service-orientation supporting the better resource utilization, increased flexibility, and enhanced provision of the user-focused environment; the governance models appropriate to driving open standards-based interoperability and integrated user services; funding models to support delivery of user-focused services leveling a cost-effective shared infrastructure provision.

The study performed allowed to formulate clear work programme for the elaboration of solutions and planning the deployment of interoperable Scientific Cloud infrastructure integrated to the national level M-Cloud initiative.

V. ACKNOWLEDGMENT

This work was supported in part by the European Commission under EU FP7 projects HP-SEE (under contract number 261499) and EGI-Inspire (under contract number 261323).

REFERENCES

- [1] “Development of Grid e-Infrastructure in South-Eastern Europe” / Antun Balaz, Ogn-jen Prnjat, Dusan Vudragovic, et al. *J. Grid Comput.* 9, 135 (2011), 22 p., DOI: 10.1007/s10723-011-9185-0.
- [2] Veaceslav Sidorenco, Petru Bogatencov, Alexei Altuhov, and Valentin Pocotilenco, “On development of National Grid Infrastructure in Moldova,” *Proceedings of the Second International Scientific Conference “Supercomputer Systems and Applications (SSA’2008).”* -Minsk, United Institute of Informatics Problems of NASB, 2008, ISBN 978-985-6744-46-7, c. 221-225.
- [3] Current state of distributed computing infrastructure deployment in Moldova / A. Altuhov, P. Bogatencov, A. Golubev, et al. “Networking in Education and Research,” *Proceedings of the 10th RoEduNet IEEE International Conference, Iasi, Romania, June 23-25, 2011, ISSN 2247-5443, pp. 69-72.*
- [4] N. Iliuha, A. Altuhov, P. Bogatencov, G. Secieru, and A. Golubev, “SEE-HP Project – Providing access to the regional High Performance Computing infrastructure,” *International Workshop on Intelligent Information Systems,* Proceedings IIS, September 13-14, 2011, Chişinău, ISBN 978-9975-4237-0-0, pp. 183-186.

On the method of sustainable use of soils by regulation of purchase orders for agricultural products

1 Eyubova Svetlana
Institute of Soil Science and Agrochemistry of ANAS,
Baku, Azerbaijan
svetlana@kiber.az

2 Pashayev Adalat,
Cybernetics Institute of ANAS,
Baku, Azerbaijan
adalat.pashayev@gmail.com

3 Sabziyev Elkhan
Kiber Ltd Company,
Baku, Azerbaijan
elkhan@kiber.az

Abstract— One of the ways to achieve of stability of soil usage is the alternation of crops. In present paper the opportunity of regulation of alternation cultivation of crops is considered as the orders for purchase of the crops.

Keywords—crop; crop rotation; purchase.

I. INTRODUCTION

Food supply is one of indicators of state's independence. In one way or another, soil is the basic source of food. In the long term, sustainable use of soils is guarantee of food safety.

Efficient crop rotation is one of sources of sustainable use of soils[1,2].

All cultivated plants produce a certain effect on soil characteristics, on inhabiting organisms, weeds, insects, microflora.

Correct crop rotation is based on environmental requirements of plants. Unlike one-crop system, correct crop rotation raises the level of crop yield. Necessity of plant rotation on the same plot of land is explained by difference in nutrition of different plants and different kinds of effect that different kinds of plants have on soil characteristics, etc. For normal vital activity, plants need numerous nutrients. That need depends on the type of plant and crop yield volume. Different plants absorb nutrients from different layers of soil due to different development of root system. For instance, grain crops have fibrous roots and use nutrients mainly from the topsoil (0-25 cm), while pulse plants and root crops, having tap roots, obtain their nutrients both from the topsoil and subsoil (>25 cm).

Plants also have different requirements to forms of inorganic nutrition. Some of them require more available phosphorus compounds, while other, such as buckwheat, can assimilate hardly soluble phosphates from the soil.

Legumes accumulate nitrogen in the soil owing to their symbiosis with root nodule bacteria. In case of long-term cultivation of leguminous crops, accumulated nitrogen cannot

be fully used by plants and even depresses further development of microorganisms and legumes themselves. Rotation of leguminous and nonleguminous crops thereby allows using all accumulated nitrogen to have heavy yields of other crops.

Crop rotation facilitates weed control. Winter crops, for instance, clear the soil from many kinds of weed. Sowing of tilled crops increases cleanliness of the soil from weeds and thereby from numerous diseases and pests.

A country's food safety is conditioned by annual state order for purchase of certain amount of agricultural produce. The list of the produce is reviewed every year and agreements with farm enterprises are made. Farm enterprises usually allocate some plots for production of concrete crops by state order.

As was said above, a number of agricultural crops are sensitive to their predecessors and their growing in the same field in a random sequence is not recommended.

It is also obvious that random redistribution of orders for crops production among different farm enterprises is fraught with another hazard; it can be economically unprofitable to fulfill the offered variant of state order. In that case, losses are usually covered at the expense of subsidies from the state budget.

Thus, we have the problem of optimal distribution of state order for crop production among all farm enterprises with further purchase of agricultural produce, with provision for crop rotation. This being the case, expenditure of each farm enterprise should be minimized to make the amount of subsidies minimal as well.

In the present paper, mathematical model is built for the process of placing of state order among farm enterprises with regard to the conditions described above.

II. THE MATHEMATICAL MODEL

Let us index all agricultural crops that can be grown in the considered region as $j = 1, 2, \dots, m$, denoting the quantity of

crops crop j specified in the plan of state order for all farm enterprises by Z_j .

Let us then index each land plot reserved by farmers for growing state-ordered agricultural crops as $i = 1, 2, \dots, n$, where n is the total number of all plots, assuming that plots $1, 2, \dots, n_1$ belong to farm enterprise N_1 , plots $n_1 + 1, n_1 + 2, \dots, n_2$ belong to farm enterprise N_2 , etc., plots $n_{k-1} + 1, n_{k-1} + 2, \dots, n_k$, ($n_k \equiv n$) belong to farm enterprise N_k , where k is the total number of all farm enterprises. Square of plot i is denoted by S_i .

Let us denote the set of agricultural crops fit for growing on plot i by \aleph_i . Assume that crop $\xi \in \aleph_i$ was grown on plot i in the previous season. It is obvious that to each such crop ξ , some subset $\aleph_i^\xi \subseteq \aleph_i$ can be assigned, which is a set of crops that can be grown on that plot, with regard to its predecessor ξ . Let us denote the set of all \aleph_i^ξ , for which $j \in \aleph_i^\xi$, by Λ_j^ξ , i.e. $i \in \Lambda_j^\xi$, if $j \in \aleph_i^\xi$.

The problem set is to select a crop for growing on allocated plots. If the decision is made to grow crop j on plot i , we will write $\delta_{ij} = 1$ and $\delta_{ij} = 0$ if not. Let us define basic relations and conditions in compliance with the problem statement.

A. Quantitative supply of state order.

Let us denote the expected productivity of crop j grown on plot i by U_{ij} . Since the aggregate amount of crop j grown on all selected plots must be Z_j with certain accuracy of admissible variation, the following equality has to hold true

$$\sum_{i=1}^n \delta_{ij} S_i U_{ij} = Z_j. \quad (1)$$

B. Condition of stable development of a farm.

When producing agricultural crops, farm enterprises carry out a number of measures, such as rent of agricultural equipment, preparation of soil for sowing, purchase and delivery of seeds, fertilizers, pesticides, seed treatment, sowing, weeding, watering, fertilizer application, etc. It is obvious that the list of measures and therefore corresponding expenses will be different for each kind of crop to be grown.

Thus, growing process is associated with different expenses for each crop j . The expenses can be different depending on the geographic location of the farm and agroclimatic conditions of the land plot. Let us denote all expenses for growing of crop j on plot i by R_{ij} . Expenses for growing of crop j in the farm enterprise N_s is calculated from the formula

$$\mathfrak{R}_j^s \equiv \sum_{i=n_s-1}^{n_s} \delta_{ij} S_i R_{ij},$$

$$(j \in \aleph_i^\xi, s = 1, 2, \dots, k, n_0 \equiv 1).$$

Therefore, to calculate expenses for the whole farm enterprise, we have the following

$$\mathfrak{R}^s \equiv \sum_{j=1}^n \mathfrak{R}_j^s = \sum_{j=1}^n \sum_{i=n_s-1}^{n_s} \delta_{ij} S_i R_{ij}. \quad (2)$$

Let us denote the purchasing price of crop j by \wp_j . Then income from fulfilling the offered variant of state order for farm enterprise N_s , ($s = 1, 2, \dots, k$) will be

$$\wp^s \equiv \sum_{i=n_s-1}^{n_s} \left(S_i \sum_{j \in \aleph_i^\xi} \delta_{ij} \wp_j \right)$$

In this case, profit of that farm enterprise is calculated by means of the following formula

$$M^s = \sum_{i=n_s-1}^{n_s} \left(S_i \sum_{j \in \aleph_i^\xi} \delta_{ij} \wp_j \right) - \sum_{i=n_s-1}^{n_s} \left(S_i \sum_{j \in \aleph_i^\xi} \delta_{ij} R_j \right).$$

Let us denote the coefficient of enterprise growth in the previous year (profit-investment relationship) by λ^s . Condition of development stability means that profit-investment relationship every fiscal year, with allowance for possible subsidies, must not be less than the respective correlation (λ^s) in the previous year. Thus, the following inequality must take place

$$M^s + D^s \geq \lambda^s \cdot \mathfrak{R}^s \quad \text{or} \\ \wp^s + D^s \geq (1 + \lambda^s) \cdot \mathfrak{R}^s, \quad (3)$$

where $D^s \geq 0$ is subsidy appropriated for farm enterprise N_s .

C. Condition of minimality of state expenses for all farm enterprises.

State order should obviously be formed in such a way that total expenses would be minimal. These expenses including amounts appropriated for purchase of agricultural crops and subsidies for individual farms, their minimality condition can be written in the following way:

$$\sum_{s=1}^k (\wp^s + D^s) \rightarrow \min. \quad (4)$$

Thus, the following problem has to be solved.

Functional (4) need to be minimized upon the condition of equalities (1) and (2), as well as inequalities (3) holding true. The problem lies in the field of linear programming, some

variables taking Boolean values. Similar problems have been sufficiently studied [3,4].

III. CONCLUSION.

Four problems can be solved simultaneously through efficient placement of orders for purchase of agricultural produce:

1. Provision of required food reserve
2. Sustainable use of soil
3. Efficiency of agricultural industry
4. Minimization of subsidy

REFERENCES

- [1] Кирюшин В.И. Экологические основы земледелия. Москва. «Колос». 1996 г. 366с.
- [2] Мовсумов З. Р. Научные основы эффективности элементов питания растений и их баланс в системе чередования культур. Баку, «Элм» 2006. 245с.
- [3] Ермольев Ю.М., Ляшко И.И., Михалевич В.С., Тюптя В.И. Математические методы исследования операций. Киев. «Вища Школа» 1979. 311с.
- [4] Романовский В.И. Алгоритмы решения экстремальных задач. Москва. «Наука» 1977. 352с.

Comparison of the Efficiency of Principal Component Analysis and Multiple Linear Regression to Determine Students' Academic Achievement

MSc. Mehtap Erguven
Computer Technologies and Engineering Faculty
International Black Sea University,
Tbilisi, Georgia
mehtaperguven@hotmail.com

Abstract:

The Georgia Ministry of Education and Science is responsible foundation to prepare the National Unified Entrance Examination (NUEE) in Georgia. Georgian Language, Logic, English Language and Mathematics are some of the categories of this examination. In this study we focused on how NUEE affects the grade point averages (GPA) of the students of International Black Sea University (IBSU). The relation between NUEE scores and GPA is represented and compared for the all students of the faculty of Computer Technologies and Engineering (CT&E) and the faculty of Business and Management (B&M). The research is also done and indicated separately for female and male students. The major purpose of this study is to compare the efficiency of multiple linear regressions (MLR) and principal component analysis (PCA) in predicting the response variable GPA using NUEE's explanatory variables (X). In the consequence, using principal components as entries improves multiple linear regression prediction by reducing complexity and high dimensionality.

Keywords: Principal Component Analysis, Multiple Linear Regressions, MATLAB Applications.

I. INTRODUCTION

What percentage of variances in graduate students' grade point average (GPA) can be explained by the national unified university entrance examination

(NUEE)? Which method is more efficient when the principal component analysis (PCA) and multiple linear regressions (MLR) were used? This article mainly focuses on these two questions. According the previous researches, prior academic successes affect university GPA [1]. This study tested the hypothesis that university GPA would be predicted by NUEE scores. There should be positive correlation between NUEE scores and university GPA, this hypothesis proved with quantitative analyses. The purpose of this study is to strengthen our prediction with usage of PCA, to compare with MLR and to improve the predictive power of MLR.

If a multivariate dataset is visualized as a set of coordinates in a high-dimensional data space (1 axis per variable), PCA can supply the user with a lower-dimensional picture, a "shadow" of this object when viewed from its most informative viewpoint [2]. Since finding a pattern is difficult and graphical illustration is not easily available in high dimensional data, PCA was our preference in this research.

II. METHODS

The data for this work was collected from the internet site of Georgia Ministry of Education and Science. <http://www.naec.ge/home.html?lang=en-GB> is the site address. The results of NUEE which had been done in 2007 were taken from this site as input data. The information of 175 students of the faculty of CT&E and the faculty of B&M was used. Analyses

were done for all students and separately for male and female students in both faculties. Our explanatory variables “predictors” are Georgian Language (GL), Logic (L), English Language (EL) and Mathematics (M) respectively. GPAs were taken from the student database of IBSU.

The MLR and PCA were performed to obtain reliable results and to compare them. All the analyses were carried out using Microsoft Excel and MATLAB. In fact, the method we used in this study can be named as principal component regression (PCR). This method combines linear regression and PCA. PCR establishes a relationship between the output variable (y) and the selected principal components (PCs) of the input variables (x_i) [3].

III. PRINCIPAL COMPONENT ANALYSIS

Principal component analysis is a variable reduction procedure. It is useful when you have obtained data on a number of variables (possibly a large number of variables), and believe that there is some redundancy in those variables. In this case, redundancy means that some of the variables are correlated with one another, possibly because they are measuring the same construct. Because of this redundancy, you believe that it should be possible to reduce the observed variables into a smaller number of principal components (artificial variables) that will account for most of the variance in the observed variables [4]. This analysis represents the contribution of each variable to the specific principal component. Instead of direct use of input variables, we change them into PCs and then we use them as input variables [5]. Our initial dataset was a set of four-dimensional (L^4) data space and we transferred it into lower-dimensional (L^2) Euclidean space.

Matrix X defined by the scalars x_{ij} is entry scores of our sample. In the matrix $X = [x_{ij}]$ the first column x_{i1} shows grades of Georgian Language, second column x_{i2} represents Logic grades, x_{i3} shows English Language marks and the last column x_{i4} indicates the grades of Mathematics of the students. Therefore the initial matrix is:

$X = [x_{ij}]; 1 \leq i \leq 175 \text{ and } 1 \leq j \leq 4$; It has four Eigen-values since the dimensions are 175×4 .

First, input data matrix X was normalized to find the PCs. The mean of each column “ \bar{X}_j ” was calculated. I used the following equation; $X_{centered} = x_{ij} - \bar{x}_j$ to normalize the matrix X. The standardized matrix is

$$X_{centered} = C = [c_{ij}] \quad 1 \leq i \leq 175 \text{ and } 1 \leq j \leq 4.$$

PCA is the Eigen-vector-based multivariate analysis. Since we use eigenvalues and eigenvectors in this analysis we can name it as Eigen-space projection. Eigen-space is calculated by identifying the eigenvectors of the covariance matrix derived from a set of facial images “vectors” [6]. Eigenvalues of the covariance matrix are used by PCA to obtain the independent axes under Gaussian assumption. Eigen values generated by PCA, also known as characteristic values, are special scalars associated with linear equation. The Eigen value for the standardized matrix is of the form

$$|K - \lambda I| = 0 \quad (1)$$

Where K is the variance-covariance matrix of the standardized data, “ λ ” is the eigenvalue and “I” is the identity matrix. In general form:

Solution is found through the Eigen problem for $Kv = \lambda v$ where $\lambda, v \neq 0$.

In our sample:

Covariance matrix of X =K ;

$$K = \begin{pmatrix} 142.8949 & 78.4146 & 38.7036 & 69.1535 \\ 78.4146 & 116.3710 & 45.5368 & 111.7319 \\ 38.7036 & 45.5368 & 99.0344 & 47.2278 \\ 69.1535 & 111.7319 & 47.2278 & 270.3662 \end{pmatrix}$$

With the covariance matrix, we can see that, all the values are positive. This means that, there is positive relationship between all pairs of dimensions. As the score of logic increased, so did the mathematics scores. These 4th and 2nd variables (dimensions) have higher relationships.

PCA was carried out to obtain the latent root ‘Eigen value’ from which the principal components were extracted; Using MATLAB:

“>> [eigenvectors, eigenvalues]=eig(cov(X));”

code gives Eigen-values and Eigen-vectors as shown below.

With the following MATLAB code, I obtained the PCs, Eigen values and Eigen vectors like previous code. The only difference is the order of the vectors and order of the Eigen-values. They are given in descending order in MATLAB:

“>>[eigenvectors, pcs, eigenvalues]=princomp(X);”
By this code, the below Eigen-values which are associated with the following Eigen-vectors are found:

$$\text{Eigen-values} = \begin{pmatrix} 390.4674 \\ 125.1986 \\ 76.2107 \\ 36.7898 \end{pmatrix}$$

According to the Eigen-values, the first principal component PC1 has the largest possible variance in this transformation. Then the second PC “PC2” has the second largest variance. By the process of taking the Eigen vectors of the covariance matrix, we have been able to extract lines that characterize the data [7].

Eigen-vectors=E;

$$E = \begin{pmatrix} -0.3956 & -0.6978 & 0.4420 & 0.4015 \\ -0.4618 & -0.2164 & 0.0261 & -0.8598 \\ -0.2470 & -0.3202 & -0.8955 & 0.1860 \\ -0.7545 & 0.6031 & 0.0454 & 0.2548 \end{pmatrix}$$

$E = [eig1 \ eig2 \ eig3 \ eig4]$; the columns of E are a set of new basis vectors which transforms C “normalized initial data” into the set of principal components P. The dot product of C and E gives matrix P. In the matrix notation, we can write the model as follows:

$$P = CE \quad (2)$$

In our study, “ c_i ” are the rows of C and “ eig_j ”s are the columns of E. C is 175x4 matrix; E is 4x4 matrix hence P is 175x4.

$$C = \begin{bmatrix} c1 \\ c2 \\ \vdots \\ cn \end{bmatrix}; n=175; \quad (3)$$

$$P = \begin{bmatrix} c1.eig1 & c1.eig2 & \dots & c1.eig4 \\ c2.eig1 & c2.eig2 & \dots & c2.eig4 \\ \vdots & \vdots & \dots & \vdots \\ cn.eig1 & cn.eig2 & \dots & cn.eig4 \end{bmatrix} \quad (4)$$

The principal components model postulates that we can summarize the variance in all X by generating an uncorrelated principal components $P = [PC1 \ PC2 \ PC3 \ PC4]$ which are a linear combinations of the original variables in X. Each column is principal component in matrix P. It is used the most significant Eigen-vector “ $eig1$ ”, to find PC1. The PC1 is 175x1 column vector. The principal components analysis search to find E such that the resulting variables in P are uncorrelated and their variances are maximized. The solution is found by using the eigenvalue decomposition of the covariance matrix

$$K = ESE' \quad (5)$$

Where S is a diagonal matrix whose diagonal elements are the eigenvalues $(\lambda_1, \lambda_2, \dots, \lambda_4)$ of K, and E is an orthogonal matrix whose columns form the set of eigenvectors. By convention, the eigenvalues are arranged in a decreasing order. Therefore, the covariance matrix for P becomes:

$$\text{var}(P) = E'KE = E'(ESE')E = S \quad (6),$$

[8]. Formally:

$$PC1 = c_{i1}*(-0.3956) + c_{i2}*(-0.4618) + c_{i3}*(-0.2470) + c_{i4}*(-0.7545); 1 \leq i \leq 175; \quad (7)$$

In the above equation, our data were in terms of the first Eigen-vector.

In the MATLAB “Xcentered*eigenvectors(:,1)” code gives the first principal component “PC1” too. The first PC1 has the highest variance of all principal components and gives the greatest amount of the variation in X.

$$PC2 = c_{i1}*(-0.6978) + c_{i2}*(-0.2164) + c_{i3}*(-0.3202) + c_{i4}*(0.6031); 1 \leq i \leq 175; \quad (8)$$

The second Eigen-vector was used to find PC2. This way helped to find all the four PCs.

“>>Xcentered*eigvectors” MATLAB code gave us the matrix P. As we expressed the PCs are perpendicular (orthogonal) to each other. To show the perpendicularity of each principal component “>>corrcoef(pcs)” MATLAB code is used. I get the unitary matrix by this code.

According the results of PCA, our study reduces the dimensionality of the transformed data with the help of first two principal components PC1 and PC2. We obtained the following estimated GPA equation:

$$\text{Estimated GPA} = 69.69583 + (-0.45537)*PC1 + (-0.11765)*PC2. \quad (9)$$

From the Eigen-values, it was clear that PC1 can explain 62% of all input data, on the other hand it was found that 82% of the total variation was explained by the PC1 and PC2 as it is seen in below figure. We can ignore third and fourth principal components. Hence, the number of principal components is reduced into two and we can still find meaningful results.

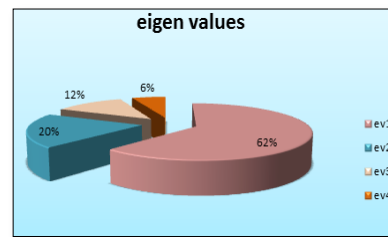


Figure 1: Total variance presentation according to the Eigen-values.

When all students were divided into two groups as male and female: For the first two PCs of male students, total variance of presentation was 0.81 and for female it was 0.83. In the B&M faculty, first two

PCs explain 89% of all male students' GPA and 82% of female students' GPA. In the faculty of the CT&E; this percentages were 80 for the male students and 79 for the female students. In all cases only two components were enough for adequate information.

IV. MULTIPLE LINEAR REGRESSIONS

MLR is one of the models to search the relationship between a dependent variable and several independent variables. In MLR, there are p explanatory variables, and the relationship between the dependent variable and the explanatory variables is represented by the following equation:

$$y = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \dots + \beta_p x_{pi} + e_i \quad (10)$$

where β_0 is the constant term and β_1 to β_p are the coefficients relating the p explanatory variables to the variables of interest and e_i is the error term. The term 'linear' is used because in multiple linear regressions we assume that y is directly related to a linear combination of the explanatory variables [9].

In our model α level is 0.05 and since the Significance F is less than 0.05 this model is significant. When we analyzed all data; the standard deviation is 16.68 "not low", the R-square is 0.26 this means that four predictors of NUÉE explains 26% of the information of the GPA. Regression analyses showed that 2 NUÉE components EL and M were significant predictors of GPA ($R^2 = .20$) and for first two variables GL and L, R-square is only 0.14. Using all variables in MLR we found the following equation:

$$\text{Estimated GPA} = -11.6549 + 0.306838 * \text{GL} + (-0.16077) * \text{L} + 0.511312 * \text{EL} + 0.373686 * \text{M} \quad (8)$$

When we compare all variables of NUÉE with GPA, the better correlation coefficient was between "logic & mathematics" and GPA. Multiple R is 0.63.

The best correlation was between the NUÉE's scores of 107 male students of the B&M faculty and their GPA. Multiple-R was 0.65 which is not so high but quite high positive correlation. This correlation for the male students of the CT&E faculty was 0.39 "low correlation", on the other side for all males multiple-R was 0.46 "low correlation".

This research indicated the correlation between the NUÉE's scores and GPA. MLR analyses results for the 68 female students showed that, the highest correlation was 0.61 for the CT&E, for the B&M faculty Multiple-R was 0.57 and for all females was 0.49 "low positive correlation".

Eventually the results of the best twenty students indicate that the students are able to have higher GPA when they start with high entrance scores. The

correlation between these students' entrance scores and GPA is higher than other students [10]. The higher entrance scores the better GPAs.

TABLE I:

Statistics		CT&E	B&M
GPA mean	male	58.6	67.3
	female	79.2	82.5
NUÉE mean	male	1644.2	1753.2
	female	1683.9	1751.1
Multiple R	male	0.29	0.48
	female	0.59	0.45

Note: Multiple R was calculated between NUÉE total scores and GPA.

In the table 1; female students of CT&E started with higher NUÉE scores and they finished with 35.2% better GPA. The correlation between NUÉE scores of female students of CT&E and their GPA is 0.59 "positive correlation". The situation was different at the B&M faculty. Male and female students almost had equal success in NUÉE but at the end of the university education female students nearly were 23% more accomplished in the faculty of B&M.

In the following figure, blue dots illustrate GPA values for 175 students. Using PCA, estimated GPA is calculated and green dots are used to show estimated GPA values without error terms. Red dots represented estimated GPA scores which were calculated by MLR. It is clear that both MLR and PCA gave very close results.

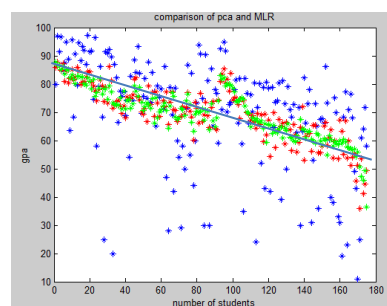


Figure 2: GPA, Estimated GPA with MLR and PCA.

CONCLUSION

The study focused on university entrance examination scores and GPAs of graduate students. It is used two scientific methods PCA and MLR to clarify the correlation between GPA and NUEE scores.

In fact, for the higher academic success in institutions, prior academic achievement measures (preparatory school grade average point (GPA), aptitude test scores, university entrance scores) and psychological variables (achievement motivation and academic self-efficacy) should be examine together[11].

Principal component analysis works on the covariance or correlation matrix to extract the directions in the multivariate space that is the “most informative”, which means, have the greatest variability. Usually, a few first components explain most of the variability in the data [12].

When the MLR analyses were implemented for all predictors of NUEE and GPA, Multiple-R is 0.51 (low correlation). When all four components were used in PCA, we had the same outcome. On the other hand correlation coefficient between “PC1, PC2” and GPA is 0.47 which is very close to 0.51. There is no need to use third and fourth PCs.

Consequently, MLR and PCA gave identical results when we use all predictors and all components respectively. Separating the data as male and female did not change the result. In this research, the regression analysis results were compared with PCA outcomes and the benefits of PCA were illustrated. This study confirms that PCA is preferable to MLR analysis in predicting the dependent variable GPA using first two components. Therefore the information of input variables was represented in PCA with minimum loses.

REFERENCES

- [1] A. Farahani, (2001), A Comparison Between Educational Self-concept and Entrance Behavior of Humanities Students in Distance and Conventional Education Systems of Iran, Payame Noor University Tehran, I.R. Iran.
- [2] Principal Component Analysis, Wikipedi, http://en.wikipedia.org/wiki/Principal_components_analysis, 2012.
- [3] A.Z. Ul-Saufie, A.S. Yahya, N.A. Ramli, “Improving Multiple Linear Regression Model Using Principal Component Analysis for Predicting PM₁₀ Concentration in Seberang Prai ,Pulau Pinang”,

International Journal of Environmental Sciences, vol 2, No 2, ISSN 0976-4402, 2011

- [4] L. Hache, “A step-by-step approach to using the SAS system for factor analysis and structural equation modeling” , SAS Publishing, ISBN: 978-1-55544-643-7.
- [5] B. Helena, R. Pardo, M. Vega, E. Barrado, J.M. Fernandez, L. Fernandez, “Temporal evolution of groundwater composition in an alluvial aquifer (Pisuerga River, Spain) by principal component analysis , 2000, Wat Res, 34: 807-16.
- [6] K. Kim, “Face recognition using principle component analysis”, University of Maryland, College Park, MD 20742, USA.
- [7] L.I. Smith, “A tutorial on principal components analysis”, February 26, 2002.
- [8] A. Grodner, W.A. Grove, “Estimating treatment effects with multiple proxies of academic aptitude”, East Carolina University, Le Moyne College, 2007
- [9] M. Tranmer, M. Elliot, “Multiple linear regression” , Cathie Marsh Center for Census and Survey Research, 2008.
- [10] M. Erguven, “Comparison of university entrance examination scores with graduate students’ GPA and effects on students’ success”, Journal of Technology and Technical Science, Vol. (1), ISSN: 2298-0032, Georgia, 2012.
- [11] O. Aboma, “Predicting first year student academic success, electronic journal of research in educational psychology”, University of Groningen, Netherlands, 7(3),1053-1072.2009(no19).ISSN:1696-2095 .
- [12] S. Kolenikova, G. Angelesb, “The use of discrete data in principal component analysis for socio-economic status evaluation”, 2005.
- [13] A.C. Kelechi, M. Okpara , American Journal of Mathematics and Statistics, “Regression and principal component analyses: a comparison using few regresses”, University of Agriculture, Umudike Abia State, , p-ISSN:2162-948X e-ISSN:2162-84752012; 2(1): 1-5doi: 10.5923/j.ajms.20120201.01, Nigeria.
- [14] R Noori, MA Abdoli, M Jalili Ghazizade, R Samieifard, “Comparison of neural network and principal component regression analysis to predict the solid waste generation in Tehran” , 2008 ,Iranian J Publ. Health, Vol.38, No.1, 2009, pp.74-84.

Computer Modeling of Hemokinins using Molecular Dynamics Method: Hemokinin 1 (human)

N.M. Qodjajev^{1,2}, B.M. Qasimov¹, U.T. Agayeva²

¹Qafqaz University

²Baku State University

Baku / Azerbaijan

Abstract – This work deals with computer modeling of human Hemokinin 1 (hHK-1) using the Molecular Dynamics Method to study the spatial structure and structural-functional relationships of the hHK-1 molecule. hHK-1 is a peptide of the tachykinin family, encoded by the TAC 4 gene. Tachykinins interact with specific membrane proteins and exhibit neurotransmitter activity.

I. INTRODUCTION

Human Hemokinin (hHK-1) is a peptide of the tachykinin family with a wide range of biological effects related to neurotransmitter activity in the interaction with specific membrane proteins.

In this paper, we study molecular dynamics of the hHK-1 polypeptide molecule. Computer modeling using molecular dynamics is an important source of information for the study of high-molecular compounds, as well as the dynamics of conformational transitions. The calculation of the "trajectory file", which contains dynamic transition between various conformations of the macromolecule, is particularly important in this respect. Trajectory file contains a set of coordinate frames, "photographed" at regular time intervals, having the order of picoseconds. Each frame holds detailed information about the coordinates of each atom of the macromolecule. Analysis of the calculated dynamical trajectories allows to build and explore the three-dimensional structure of the molecule at each time moment, define the length and angular characteristics of the various types of bonds between atoms and groups of atoms that form the molecule.

II. DESCRIPTION OF THE MODEL

In the numerical molecular dynamic simulation method the system is represented as a set of material points, the interaction of which is described by the given force fields, and the dynamics are described by the classical equations of Newton.

The primary structure of the molecule is shown in Fig. 1.

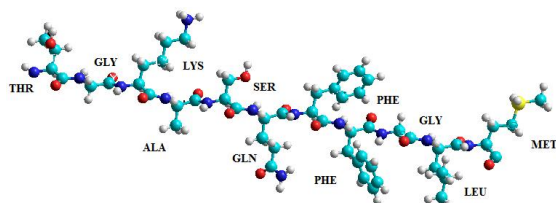


Fig.1. The primary structure of the hHK-1 molecule.

The polypeptide structure, shown in Figure 1 was built using the "HyperChem" Molecular Modeling System [1], according to the following sequence of the anions:

(+THR-GLY-LYS-ALA-SER-GLN-PHE-PHE-GLY-LEU-MET-z).

The calculation was performed using the GROMACS software package [2] for molecular dynamics simulations of biomolecular systems. As a result of the calculation the dynamic trajectory of the system, consisting of the molecule in an aquatic environment, was obtained for 10 ns, using the "GROMAS" force field; the thermostat temperature was set to room temperature (300°K). The system cell in an aqueous medium with a water model SPC is shown in Fig. 2.

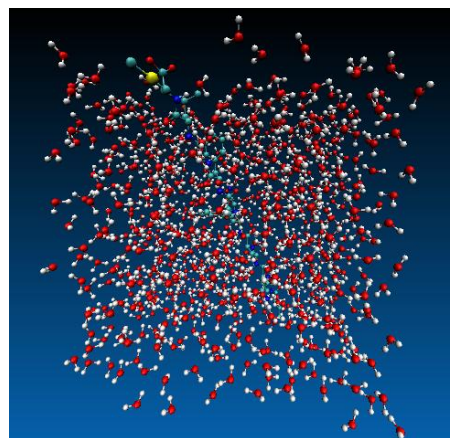


Fig. 2. Working cell: molecule of hHK-1 in aquatic environment.

II. THE CALCULATION RESULTS

The dynamic trajectory of the macromolecule was obtained as a result of the conducted calculations. The obtained trajectory was used then to analyze the spatial structure of the molecule in the process of conformational transitions and to define the secondary structure of the molecule. This trajectory was used also to calculate the distribution of the prime and dihedral angles between the bonds of various groups of atoms.

The spatial arrangement of the molecule for a fixed time frame is shown as an example in Figure 3.

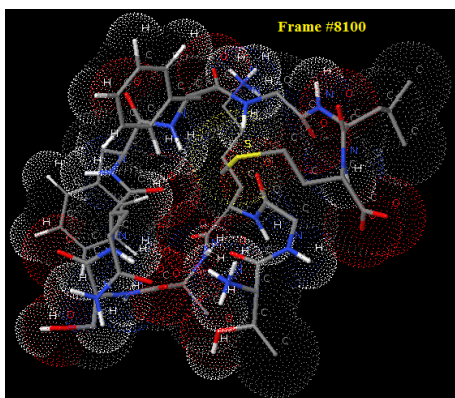


Fig. 3. Arrangement of the hHK-1 molecule in the 3D space.

The distribution of the angles ϕ and ψ for the various constituent components of the polypeptide chain (Ramachandran map) is presented in the Figure 4 for the same time frame.

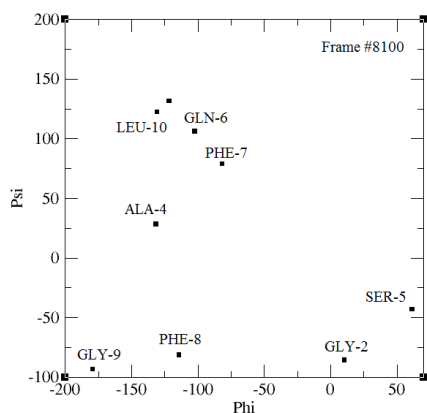


Fig. 4. Ramachandran map for the ϕ and ψ angles at the same fixed time moment.

The calculated trajectory files were also used to determine the secondary structure of the molecule. The parameters of the secondary structure were calculated basing on atomic coordinates in the molecule by the method of recognition using an informational database [3]. In the table below the resulted secondary structure for the time frame used earlier is presented:

Residue	Structure	Phi	Psi	Area
THR - 1	T Turn	360.00	129.28	110.4
GLY - 2	T Turn	10.16	-85.59	15.7
LYS - 3	T Turn	-131.70	28.46	145.4
ALA - 4	T Turn	-121.72	131.83	46.2
SER - 5	T Turn	61.38	-42.79	97.8
GLN - 6	T Turn	-102.64	106.19	83.6
PHE - 7	T Turn	-81.92	79.14	155.1
PHE - 8	C Coil	-114.22	-81.28	142.6
GLY - 9	C Coil	-179.53	-93.05	30.4
LEU - 10	C Coil	-130.86	122.72	169.8
MET - 11	C Coil	-147.24	360.00	158.8

The dynamic trajectory of the molecule was used to analyze the distribution of dihedral angles for each residue separately, and for the molecule as a whole.

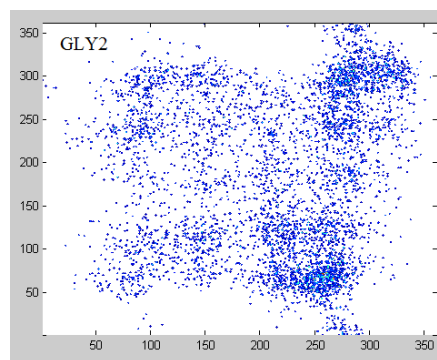


Fig. 5. Distribution of the angles ϕ and ψ for the GLY2 residue taking into account all conformations during the whole trajectory.

The result of the calculation of the ϕ - ψ distribution is presented in the Figure 5 in the form of contour plots for the GLY2 residue. For comparison, the probability distribution of angles for all GLI residues in the hHK-1 molecule taking into account all their conformations throughout the dynamic trajectory is given in the Figure 6.

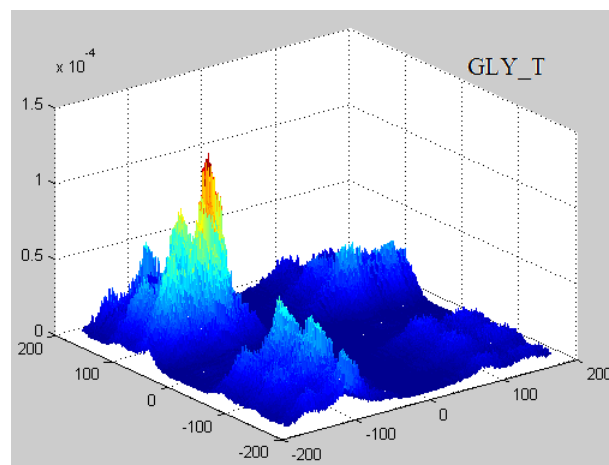


Fig. 6. The probability distribution of angles ϕ and ψ in the form of a histogram for for all GLY residues during the trajectory.

The calculations described here were made using a force field GROMOS-96 [4]. In the following we propose to repeat all the calculations using the force field AMBER-99 [5] and conduct a comparative analysis of the results.

ЛИТЕРАТУРА

1. HyperChemTM, Molecular Modeling System, © 2007 Hypercube, Inc.
2. B. Hess and C. Kutzner and D. van der Spoel and E. Lindahl GROMACS4: Algorithms for highly efficient,

- load-balanced, and scalable molecular simulation; *J. Chem. Theory Comput.* 4, pp. 435-447, 2008.
3. Frishman D, Argos P. Knowledge-based protein secondary structure assignment; *Proteins* 23(4), pp.566-79, 1995.
 4. Van Gunsteren, W. F., Billeter, S. R., Eising, A. A., Hunenberger, P. H., Kruger, P., Mark, A. E., Scott, W. R. P., Tironi, I. G. *Biomolecular Simulation: The GROMOS96 manual and user guide*. Zurich, Switzerland: Hochschulverlag AG an der ETH Zurich, 1996.
 5. Wang, J., Cieplak, P., Kollman, P. A. How Well Does a Restrained Electrostatic Potential (RESP) Model Perform in Calculating Conformational Energies of Organic and Biological Molecules. *J. Comp. Chem.* 21(12), pp. 1049–1074, 2000.

Using of Information and Communication Technologies in Learning and Teaching of Physics in Universities: Few Parametric Model for the Dynamics of Vibrations of Diatomic Molecules

R. Qadmaliyev¹, B.M. Qasimov¹, N.M. Qodjayev^{1,2}

¹Qafqaz University
²Baku State University
Baku / Azerbaijan

Abstract – Using of Information and Communication Technologies in Education both for learning and teaching of Physics is very important component in university education today. The new approach used in Qafqaz University for learning and teaching of Physics was reported in our recent papers [1,2]. The essence of the approach is based on interactive numerical modeling in Physics. In this article we continue developing this approach considering the unified numerical model for the dynamics of diatomic molecules.

Keywords - ICT in Education; active and interactive learning; molecular dynamics, vibrational dynamics, normal modes, visualization

I. INTRODUCTION

In recent papers [1,2], we reported on our approach to the use of computer technologies in Qafqaz University as a medium for teaching and learning physics. The essence of the approach is based on interactive numerical modeling in Physics. In this paper this approach is illustrated by considering the numerical model of diatomic molecules, using simple analytical potential with a small number of parameters, which allows to investigate both the molecular dynamics of diatomic molecules and the dynamics of their vibrations.

Many studies of molecular systems, undertaken "in silico", now are conducted using ready software packages that are realization of classical molecular dynamics and modern methods of quantum mechanics [3]. The use of such software packages greatly facilitates the work of researchers, but on the other hand the details of the algorithms can remain dark. In this connection it is appropriate from a methodological point of view during the teaching process at the Physics Department of the University, as part of the course "Computer simulations in physics," to give students the opportunity to model the dynamics of simple molecules using analytical potential function with a minimum number of parameters. In this case, the same numerical model of a molecule can be used to

calculate a dynamic trajectory of the molecule, and the frequencies of the normal vibration modes.

In the development of this educational model, and visualize the results of the calculation we used the Borland Builder C++ v.6 programming environment [4].

II. DESCRIPTION OF THE MODEL

From algorithmic point of view the model is based on the concept of the normal vibration modes of the molecule, which in turn is initially contained in the molecular dynamics model, built on the basis of the Newton's classical equation of motion:

$$\vec{F} = m \frac{d^2 \vec{r}(t)}{dt^2} \quad (1)$$

The forces acting on the atoms are determined by the potential:

$$\vec{F} = -\nabla U \quad (2)$$

In our case, the potential is a function of the coordinates of the atoms and defines the various components of this interaction, which are included in the consideration.

Solution of the equation (1) in molecular dynamics method usually is conducted by the Verlet integration [6]:

$$\begin{cases} \vec{v}(t + \frac{\Delta t}{2}) = \vec{v}(t - \frac{\Delta t}{2}) + \frac{\vec{F}(t)}{m} \Delta t \\ \vec{r}(t + \Delta t) = \vec{r}(t) + \vec{v}(t + \frac{\Delta t}{2}) \Delta t \end{cases} \quad (5)$$

Here Δt is the time increment.

Then the values of acceleration for each nuclear center, accounting their masses, and the forces acting on atoms are calculated. From this data, the system of equations of motion is solved and the new coordinates of the atomic centers are

calculated, which in turn are used for visualization of molecular movement and calculations on the next iteration step.

We want students to study the role of pair and many-body interactions in the molecule, which are important in the formation of the vibrational spectrum of the molecule. In general form the potential energy of the molecule can be represented as the sum of contributions of various interactions between atoms. In the simplest case, when considering the CO₂ molecule, we restrict ourselves to the two-body interaction, using a 3-parameter analytical Morse potential [5] and noncentral many-particle interaction in the form of a parabolic potential, which depends on the bond angle:

$$U_1(\mathbf{r}) = p_0 \cdot e^{-2 \cdot p_1 \cdot (r - p_2)} \quad (3)$$

$$U_2(\alpha) = \frac{1}{2} p_3 (\alpha - p_4)^2 \quad (4)$$

In the harmonic approximation the normal modes of the molecule are calculated using the force constants matrix, which in general is of dimension $N \times N$ and is defined by the second derivatives of the total potential with respect to the local coordinates of the atoms. The expression for the force matrix is given below:

$$k_{ij} = \frac{\partial^2 U}{\partial x_i \partial x_j}, \quad i, j = 1, 2, \dots, N$$

In order to calculate the normal vibration modes on the basis of the interaction potential, one has to construct the dynamical matrix, using (5) and taking into account the mass coefficients:

$$\tilde{k}_{ij} = \frac{k_{ij}}{\sqrt{m_i} \sqrt{m_j}} \quad (6)$$

Then one has to solve the eigenvalue problem, using, for example, the Jacobi method [8]:

$$(\tilde{k}_{xx}^{ij} - \omega^2 \delta_{ij}) \tilde{x}_j = 0 \quad (6)$$

Here \tilde{k}_{xx}^{ij} is the Hessian, constructed along the axis of the linear molecule.

III. USING THE MODEL IN EDUCATION

On the basis of the algorithm described above we have designed an interactive program that allows students actively investigate the normal modes of a diatomic molecule. Working with the program students can specify components of potential, which is responsible for the interaction of atoms in the

molecule and, thus, to investigate the influence of the role of non-central interactions in the formation of the vibrational spectrum of the molecule. The program has two main functioning modes - one for the calculation of normal vibration frequencies, the other - for the visualization of molecular vibrations using molecular dynamics.

Interface of the program in both modes, and the interactive windows, which are used to change parameters of the problem are shown in the figures below. The Figure 1 shows the window in the visualization mode, demonstrating the vibrational dynamics of the CO₂ molecule.

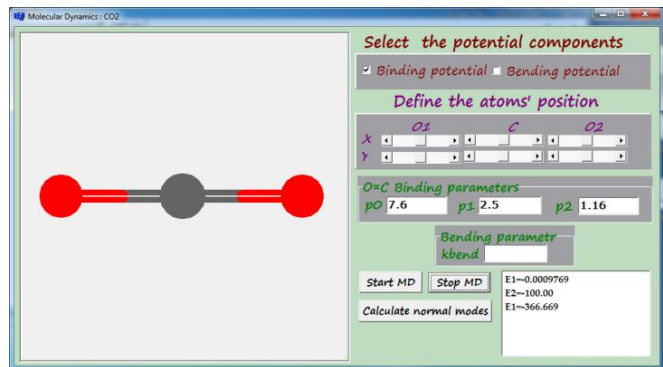


Fig. 1. The interface of the program visualization mode for the described model in C++ Borland Builder 6 environment.

In the visualization mode students can interactively include into the consideration the potential component (4), which is responsible for the nonpair interactions of atoms. Changing of the initial local coordinates of the atoms in a molecule can be provided from this window, using appropriate scroll bars, which are available for each atom.

The next figure illustrates another screenshot of this form at the time of animation including asymmetric vibrational mode of the molecule.

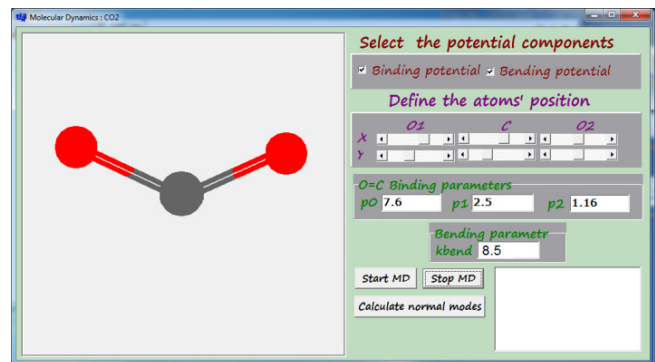


Fig. 2. The visualization of the molecule dynamics of the CO₂ molecule.

In the Figure 3 the interface of the second mode window form for the calculation of the frequencies of the normal vibrations of the CO₂ molecule is shown.

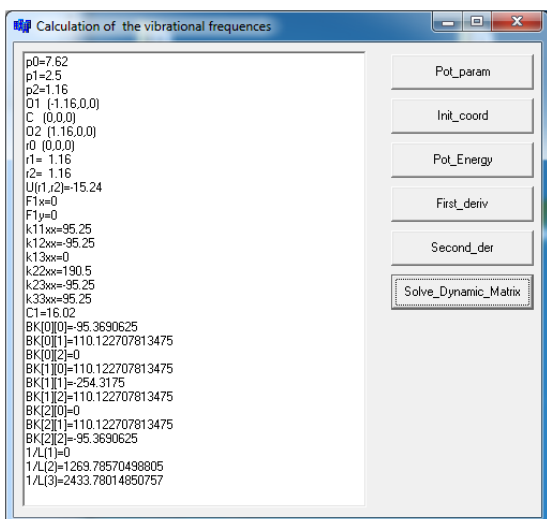


Fig. 3. Interface of the program realisation in the normal frequencies calculation mode.

As can be seen in Figure 3, the design of the form sets the sequence of steps to calculate the frequencies of the normal vibrations of the molecule. First of all, a student identifies potential parameters (Button "Pot.Param."); Then - initial local coordinates of the atoms (Button "Init. Coord."); Then, if needed, can calculate the potential energy of a given conformation (Button "Pot.Energy") and see if in the equilibrium structure appropriate to the minimum of potential energy; It is possible to calculate the first derivative of the potential of the current conformation (Button "First_deriv.") and the local curvature of the potential surface (Button "Second_der."); finally, students can solve the complete eigenvalue problem for the Hessian and determine the normal frequencies (Button "Solve_Dynamic_Matrix").

In the normal frequencies calculation mode it is convenient to use separate dialog boxes to set parameters of the potential and initial values of atomic coordinates.

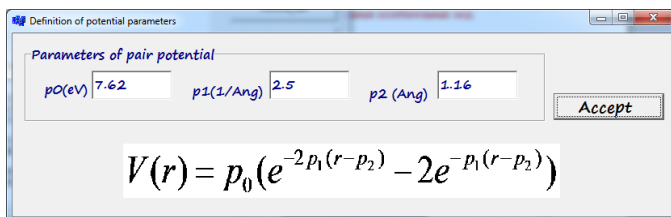


Fig. 2. The interface of the interactive window for defining the parameters of the two-particle Morse potential.

In the figure below the interactive form for defining the initial local coordinates of the atoms in a molecule is shown. Depending on the values of these local coordinates students can watch various modes of normal vibrations of the molecule.

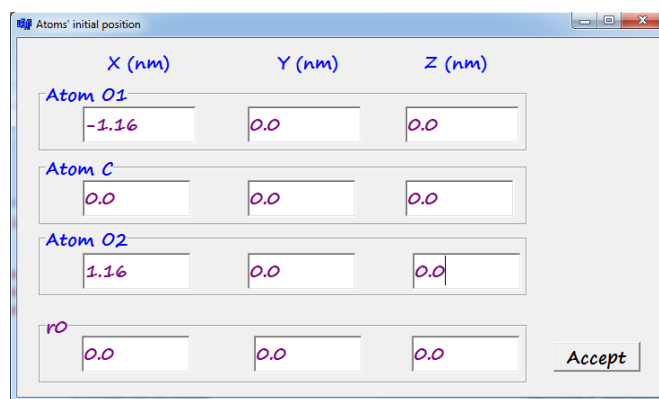


Fig. 3. The interface of the interactive window for entering atoms' local coordinates.

Using this mode students can arrange active research of the dependence between the normal frequencies and parameters of the used potentials. Visualization of the potential itself can be provided, for example, using MATLAB [7]. The pair potential for the CO₂ molecule considered in this model is shown in the figure below.

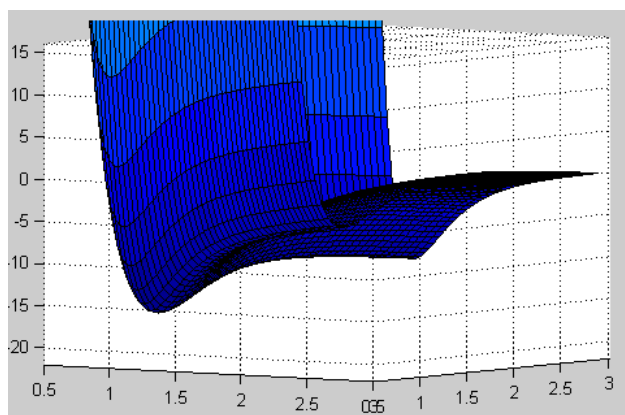


Fig. 2. The pair potential for CO₂ molecule.

We choose a CO₂ molecule because of the fact that the molecule has a simple structure and a linear symmetry, making it possible to confine with system of equations having the 3rd order, when calculating the eigenvalues of the Hessian, describing the vibrations along the molecule axis. In this case, the availability of data on the IR spectra of this molecule make it possible to compare calculated frequencies with the experimental values of symmetric and antisymmetric vibrational modes.

Including many-particle interactions into the consideration, represented by the formula (4), allows students to investigate the role of non-central interactions when comparing the calculated frequencies of normal modes with experimental values.

REFERENCES

1. N.M.Qodjajev, B.M.Qasimov. Using of Information and Communication Technologies in Learning and Teaching of Physics in Universities. - Conference Proceedings of the 3rd International Conference on Application of Information and Communication Technologies, IEEE, Baku, October 12-14, 2009, p.82-84.
2. N.M.Qodjajev, B.M.Qasimov. Using of Information and Communication Technologies in Learning and Teaching of Physics in Universities: Unified Interactive Numerical Model for Two Body Problem.- 5th International Conference on Application of Information and Communication Technologies, IEEE, Baku, October 12-14, 2010.
3. "Journal of Computational Chemistry", 2005, v. 26, n. 16.
4. Borland C++ Builder, Version 6.0, 1983-2002 Borland Software Corporation.
5. И.Г.Каплан, Введение в теорию межмолекулярных взаимодействий, Москва, Наука, 1982.
6. M.P.Allen and D.J.Tildesley, Computer Simulation of Liquids, Oxford, Clarendon Press, 2002.
7. Matlab R20010b, MathWorks, 1984-2010.
8. Н.Н.Калиткин, Численные методы, Москва, Наука, 1978.

TEACHER'S ROLE IN THE PROCESS OF CALL IN LANGUAGE TEACHING

Ali Shahintash
Qafqaz University,
Department of Computer Engineering
Baku/Azerbaijan
asahintas@qu.edu.az

Mehmet Shahiner
International Black Sea University,
Faculty of Education & School of Languages
Tbilisi/Georgia
msahiner@ibsu.edu.ge

Abstract - Computer assisted language learning (CALL) refers to the sets of instructions, which need to be loaded into the computer for it to be able to work in the language classroom. It should be borne in mind that CALL does not refer to the use of a computer by a teacher to type out a worksheet or a class list or preparing his/her own teaching alone. Teacher's participation and roles should be taken into consideration to get good impact in language teaching while using computer.

Keywords: education, CALL, teacher's role, language teaching

INTRODUCTION

Some researchers emphasized that the importance of teacher presence in the classes with their comments. What is more, some research studies indicated the essential roles of teachers in CALL in language teaching.

Integrating computers into language classes is a curriculum change, and attempts taken for this change require the involvement of teachers in all stages like preparation, forming, resourcing, promoting and implementation, so who is to implement this change in this context is quite apparent: teachers as being the part of the process.

Therefore, teachers' inquiries and reflections should be taken into consideration to avoid from any failure throughout the application of innovations (Chapelle).

As Crook (1994) stated, *If we do wish to conduct evaluations of what is learned in computer-based contexts, we must go beyond the input-output designs that characterize much research in the area... Computers are unlikely to function as magic bullets effortlessly releasing their therapeutic effects at points identified by the teachers. The unfamiliarity and wizardry that surrounds them may cultivate such notions, but the real impact of learning through this technology may need to be measured with attention to how it is assimilated into the surrounding frame of educational activity (Ediger).*

Meskill, Mossop and Bates (1999) mentioned that a person modeling the behavior or task and observing the students' learning process can be helpful for them to develop internal control. Similarly; Horwitz, Horwitz and Cope (4) stated that particularly in a second language learning environment students need much more guidance and to provide guidance is absolutely the teacher's responsibility

TEACHER'S ROLE IN THE PROCESS OF CALL IN EDUCATION

CALL is an immensely important resource for education, learners and teachers as well, it presents them genuine opportunities for self-controlled learning, the complete wealth of materials available present more roles on the educator, (Pienemann, 1984). CALL also motivates students to develop extra learning objectives as well as agendas. Yet, owing to CALL's structure; there is a high possibility for confusion: students, particularly those with lower ability may not fully develop navigational proficiency required to obtain what they want and techniques to utilize materials. For the analysis of this discourse, teachers' role remains important.

CALL would not take place without language teachers; these teachers as well as learners have to be interested as well as enthusiastic. However, regardless of the present established availability of CD-ROMS and web essential for grammar teaching, it is by no chance obvious that the technique has gained popularity and assurance estimated (Warschauer, 1995). The attitude of the teachers (those supervising learners in computer exploitation) seems to be more complex. Majority of grammar instructors (especially, the elder generation) are not comfortable with sophisticated technology, because unlike many learners, they grew-up without computers. As such, they fear that they can be replaced through introduction of computers; others dislike the idea of spending huge sums of money on sophisticated technology instead of, for example, on books, classrooms and

so on. In any grammar center a certain proportion of teachers will remain antagonistic to or basically uninterested in CALL. However, there is slight uncertainty that as innovation has grown less multifarious, recognition as well as importance among teachers have become further predominant. It is imperative to note that what prevents teachers to fulfill CALL process is inadequate time, because they seem to be adequately loaded by their traditional administrative class responsibilities.

In addition, they acknowledge that computers do not eventually save or minimize workload, as they are expected. Even though they were offered technical and pedagogical training, still they won't find sufficient time to practice what they have learnt. Therefore, CALL might be minimized to application, where educators only manage to propose to their learners CD-ROMs or websites significant to the curriculum, lacking adequate time to appropriately incorporate CALL into regular teaching, or assist students independently by offering exercises in self-access manner (Levy, M. 1997). In this case, not surprisingly, CALL application will be characterized by low efficiency.

Computers, as we have already shown, do not free language teachers of any duties, they just optimize their time (instead of writing the same things on the blackboard they can use power-point presentations) to let them participate in more communicative activities with students.

Computers should not be perceived by language teachers as competitors as nothing can substitute natural face-to-face language communication. They should be perceived by language teachers as devices that permit to unload them from mechanical work and making their work more creative.

The main reason for the failure of CALL to accomplish its pledges was the inadequate participation of grammar teachers. Additionally, there is a pivotal question: Are students interested in the benefits of CALL? These issues have received little concentration on CALL studies; as such it requires to be solved if the technique has to be effective in English language teaching (Warschauer, 1995a).

Grammar teaching entails several activities ranging from presentation of precise input to interaction output of a variety of communicative exercises. It is evident that language teacher's performance in these numerous sorts of activities is diverse. In presentation, for instance, the teacher is 'strong in direction' while in interaction output the 'task of the instructor should shift' (Harmer, J.). Nevertheless, the role of teacher in CALL depends on the roles they conduct in various exercises. The success of CALL in grammar teaching depends wholly on instructors, whereas it may be contributing ultimately to self-sufficiency, it can't be considered as a basically self-access process. Certainly, it requires student teaching and supervision to be conducted by the instructor.

Thus, how can grammar teachers manage the application of computer in the process of language learning? Obviously, CALL involves various degrees of instructor involvement (Ellis, R. 1994). In a well-developed and interactive grammar teaching students may feel largely controlled or totally free. Subsequently, in any CALL exercise at any stage of

proficiency, for several cultural or emotional reasons, some students will have more educator-reliance. A CALL exercise, for example, that comprises of a small team or entire class requires an educator to be consistently on hand to provide guidance.

The first question that arises is where is CALL used: in the classroom or in the process of individual study in computer lab or at home? While students are not very experienced in the application of computer as a learning tool, holding classes partially or totally dedicated to CALL are a necessity. When CALL is used in the classroom, the teacher's role is:

- to decide (taking into consideration the number of computers in the classroom, on the one hand, and peculiarity of the task, on the other) whether students will work individually, in pairs or in small groups, simultaneously or in a certain order, and to organize students accordingly

- to decide (taking into consideration students' age, their computer and language skills' level) whether the computer will be used as an information source, as a tutor (to fulfill drills with feedback), as a communication means (e-mailing, conferencing) or as an assessor of grammar skills and to instruct the students on the task

- to recommend which software to use, taking into consideration students' individual peculiarities

- to assist students in case of some technical or linguistic problems when/if they arise

- to monitor the process (to control that students are really doing the task) and provide the discipline

- to organize the summing up stage (students making presentations; fulfillment of traditional drills and activities not based on the application of computer that permit to see the results of CALL activities; assessment)

If the last stage does not follow, using computer in the classroom is just a waste of time and the teacher may be even blamed of having a rest and not performing his/her professional duties.

When CALL is taking place out of classroom (computer lab, internet café, home), teacher may (if it is given as home task) or may not (if CALL is realized on student's initiative) manage the process. More often it is done individually, however, work may be done in pairs or small groups. Teacher may monitor the work on students' request. Again, as in classroom application of CALL, the final stage of presenting results in the classroom is essential.

Kern (1996) notes that a shift from the use of the computer for drill and tutorial purposes to a medium for extending education beyond the classroom and reorganizing instruction has resulted in role changes for both learners and teachers. Learners now view the computer as a medium through which they must negotiate meaning through interaction, interpretation, and collaboration rather than as a finite, authoritative informational base for carrying out a stipulated language task. Instead of delegating language instruction to the computer, teachers participate in students' communication and learning and "provide a scaffold for their students' learning with their own knowledge and experience - even

when they are not immediately involved in a communicative exchange".

Reports in the research note that teachers' jobs are harder in the early stages of a technology's implementation, that positive changes from technology are more evolutionary than revolutionary, and that these changes occur as teachers become more experienced with the technology (Weiss, 1994).

REFERENCES

- [1] Chapelle, C. A. (2001). Computer applications in second language acquisition. Cambridge : Cambridge University Press. / (2003). English language learning and technology: Lectures on applied linguistics in the age of information and communication technology. Philadelphia: John Benjamins.
- [2] Ediger, M. Elementary education. Kirksville, Missouri. Simpson Publishing Company. p. 47-60
- [3] Crook, C. (1994). Computers and the collaborative experience of
- [4] Meskill, C., Mossop, J., & Bates, R. (1999). Electronic texts and English as a second language environments. New York: National Research Center on English Learning and Achievement. (CELA)
- [5] Horwitz, E. K., Horwitz, M. B., & Cope, J. A. (1991). Foreign language classroom anxiety. In E. K. Horwitz, & D. J. Young (Ed.). Language anxiety (p. 27-39). Englewood Cliffs, NJ: Prentice Hall.
- [6] Pienemann, M. 1984. Psychological constraints on the teachability of languages. *Studies in Second Language Acquisition*. 6(3): 186-214.
- [7] Warschauer, M. (1995a). E-Mail for English teaching. Alexandria, VA: TESOL Publications. (2/3)
- [8] Levy, M. (1997). CALL: Context and Conceptualization. Oxford: Oxford University Press.
- [9] Warschauer, M. (1995a). E-Mail for English teaching. Alexandria, VA: TESOL Publications. (2/3)
- [10] Harmer, J. How to teach English
- [11] Ellis, R. 1994. The study of second language acquisition. Oxford: Oxford University Press.
- [12] Kern, R. (1996). Computer-mediated communication: Using E-mail exchanges to explore personal histories in two cultures. In M. Warschauer (Ed.), *Telecollaboration in foreign language learning: Proceedings of the Hawai'i symposium* (pp. 105-109). Honolulu: University of Hawai'i, Second Language Teaching & Curriculum Center.
- [13] Weiss, J. (1994). Keeping up with the research. *Technology and Learning*, 14(5), 30-34.
- [14] Willoughby, S.E. (1993). Card Game Activities Using Grammar-Based Dialogues. MA thesis. School for International Training, Brattleboro, USA
- [15] Yaman, H. (2010). Cartoons as a Teaching Tool: A Research on Turkish Language Grammar Teaching. *Educational Sciences: Theory and Practice*. p.1231-1242.
- [16] Rose, Kenneth H. (2005). *Project Quality Management: Why, What and How*. Fort Lauderdale, Florida: J. Ross Publishing. p. 41.
- [17] Oxenhandler, N. (1988). The changing concept of literary emotion. *A selective history*. *New Literary History*, 20, p. 105-121.
- [18] Ormond, J.E. (2008). *Educational Psychology: Developing Learners*. Merrill. Pearson Education. AllynBakon, Prentice Hall learning. London: Routledge.
- [19] Krantz, J. A. (2009). Teaching Grammar for Communication in the Secondary French Classroom. In McCoy, L.P. (ed.) *Studies in Teaching: 2009 Research Digest*. Research Projects Presented at Annual Research Forum (15th, Winston-Salem, North Carolina, December) p. 55-61
- [20] Kauchak, D. and Eggen, P. (2005). *Introduction to Teaching*. Pearson Education International.
- [21] Jurjill, D.A. (2011). Propelling Students Into Active Grammar Participation. University of Applied Sciences Utrecht. Report retrieved December 12, 2011 from <http://www.eric.ed.gov/PDFS/ED521059.pdf>
- [22] Jonassen, D. (1999). Designing constructivist learning environments. In C. M. Reigeluth (Ed.), *Instructional design theories and models. A new paradigm of instructional theory* (pp. 215-239). Mahwah, NJ: Erlbaum.
- [23] Astleitner, H. (2004). Multimedia Elements and Emotional Processes. *E-Journal of Instructional Science and Technology*, v7. n2. Retrieved December 12, 2011 from <http://www.eric.ed.gov/PDFS/EJ850353.pdf>

AUTHOR INDEX

A

A. V., Maslov	325
Abbasli, R.M.	389, 438
Acar, Yusuf	181
Adalat, Pashayev	504
Adamov, Abzetdin	125
Afzali, Shima	3
Agaeva, G.A.	402, 461
Agaeva, L.N.	438
Agayeva, U.T.	402, 512
Agrawal, Arjun	57, 108
Ahola, Jukka	23
Aitpayev, Kairat	95
Akcayol, M. Ali	441
Akhmedov, N.A.	389, 438
Akhmedova, S.R.	389
Akverdieva, G.A.	465
Alhusaini, Lina	393
Ali Attea, Bara'a	441
Aliyev, Alovzat Q.	474, 477
Aliyev, K.A.	347, 371
Alizadeh, Farhad	253
Alkhatib, Bassel	421
Alnahhas, Ammar	421
Alomari, Ahmad	231
Ar, Yilmaz	273
Archuadze, M.	417
Arora, Rohan	198
Aslanov, A.A.	405
Ayob, Adenan	381, 495

B

Badura, Stefan	35
Bahri, Mammad	340
Bakhishova, Vusale	340
Balogh, Zoltán	450
Bardavelidze, Avtandil	470
Bardavelidze, Khatuna	470
Basiladze, George	308
Bayram Cetin, Gulsah	119
Bayramova, Nargiz	367
Besiashvili, G.	417
Bogatencov, P.	499
Bora, Margarita	165
Brusbardis, Valters	138

C

Celik, Elif Tuba	143
Cha, Suk Won	268
Chaudhry, Harish	80
Chechel, A.O.	426
Chikovani, Davit	19
Cooklev, Todor	181
Cosar, Ahmet	12
Costanzo, Alfio	30

D

Darbandi, Mehdi	7
Datbayev, Zhanibek	128
Deghani, Hamid	429
Digulescu, Angela	143
Dilek, Selma	3
Doğan, Hakan	181
Dokeroglu, Tansel	12
Dong, Jian	113, 358
Drlik, Martin	236
Du, BoWen	113, 358
Dvalishvili, P.	344

E

Elbası, Ersin	242
Elkhan, Sabziyev	504
Elsts, Atis	248
Erguven, Mehtap	507
Eshet, Yovav	485

F

Fallahpour, Mojtaba Behzad	429
Farhangian, Nooshin	39, 192
Faro, Alberto	30
Feizi-Derakhshi, Mohammad-Reza	263
Fekr, Nasim	192
Firoozy-Najafabadi, Hamid-Reza	75, 133, 263

G

Gaber, Jaafar	95
Gachechiladze, Lia	19
Gasimli, A.N.	371
Ghlonti, Giorgi	69

Ghvaberidze, Bezhan	313
Girgyliani, Akaki.....	333
Gochitashvili, Ketevan	364
Godjaev, N.M.....	402
Godjayev, N.M.....	438
Granmo, Ole-Christoffer	7
Grinautski, Keren	485

H

Habibizad Navin, Ahmad	133
Harahap, Erwin.....	207
Hashemi, Seyyed Mohsen	328
Hilal, Saba	105
Huang, Jian.....	113, 358
Humbetov, Shamil.....	147
Hurezeanu, Bogdan	143
Huseynov, S.T.....	405

I

Ibadov, N.V.....	405
Iliuha, N.....	499
Ismailova, L.I.	389, 438

J

Jahangiri, Nesa	3
Jain, Darshan	198
Jiang, Zetao	410

K

Kabulov, B.T.....	62
Kalagiakos, Panagiotis	165
Kamyab Hesari, Manoochehr.....	253
Kang, Jianchu	358
Kapanadze, Mikheil	299
Kapusta, Jozef	236
Karacan, Hacer	3
Karasha, Toms.....	155
Karayev, R.A.....	347, 371
Karbalae, Hassan	53, 171, 258
Kazimova, N.E.	347
Kervalishvili, P.....	417
Khachidze, M.	417
Khakmardan, Samaneh.....	353, 481
Khan, Ahmad	198
Khanna, Seema.....	80
Khosravi, Hasan	481

Khutsishvili, Irina	313
Klimo, Martin	35
Koprda, Štefan	450
Kosari, Elyas Mohamadzadeh.....	285
Kozlinskis, Emils	138
Krak, Iurii.....	414
Krishna A., Vijaya	159
Kryvonos, Iurii.....	414
Ksouri, Moufida.....	376

L

Laabidi, Kaouther	376
Leonardi, Rosalia	433
Lim, V.G.....	62
Lin, Yufei.....	217
Liu, Lieli	203
Liu, Tao.....	203

M

Mahamatov, Nurilla	268
Maiorana, Francesco	433
Makhambetov, Olzhas	128
Maklaev, V.....	64
Maleki Javan, Alireza	253
Mamaghani, Ali Safari.....	489
Masiyev, Khayyam H.	340, 367
Mazilu, Madalina	143
Md. Hashim, Aslinda	88
Mednis, Artis.....	248
Meybodi, Mohammad Reza.....	489
Midodashvili, B.....	344
Mohamadian, Habib.....	455
Mohammed, Abbas	212
Mostofi, Mehdi	385
Mshvidobadze, Tinatin	185
Munk, Michal.....	236
Muradkhanli, Leyla.....	447

N

Nagiev, M.A.....	347
Narwal, Neetu	105
Navin, Ahmad Habibizad.....	75
Nejad, Mehrpooya Ahmadali	328
Nejati, Omid.....	71
Nishi, Hiroaki.....	207
Nouri, Mahdi.....	39, 192

O

Okujava, Shorena	318
Osman, Zosipha Zainal	88
Özdemir, Suat.....	242, 273, 398, 441

P

Pang, Songsong	113, 358
Paposhvili, Mariam	336
Payne, A.I.L.	371
Peled, Yehuda.....	485
Popescu, Theodor D.	221

Q

Qadmaliyev, R.....	515
Qasimov, B.M.	512, 515
Qasymov, Ilkin.....	340
Qodjayev, N.M.....	512, 515

R

Raj, Nishant.....	198
Riahi, Aref.....	353, 481
Rodonaia, Irakli.....	45
Rodonaia, Vakhtang.....	45, 49
Romli, Fakaruddin Fahmi	88
Rustamov, Anar.....	176

S

Sadikhova, N.Y.	371
Salehifar, Mohammad Reza	53, 171, 258
Salih, Sami H. O.....	212
Samkharadze, Roman.....	19
Savola, Reijo M.....	23
Secrieru, G.....	499
Selavo, Leo.....	248
Setayesh, Saeed	7
Seth, Richa	226
Shahbazi, Pariya.....	7
Shahiner, Mehmet	519
Shahintash, Ali	519
Sharma, Priyanka	57, 108
Shen, Bin	128
Shrivastava, Dhruvad	198, 226
Singh Bindra, Gundeep	57, 108, 198, 226
Siraczade, Elvira.....	367

Sirbiladze, Gia.....	299, 313
Skliarova, Iouliia.....	291
Sklyarov, Valery	291
Skvarek, Ondrej	35
Soleimany, Saeed.....	53, 171, 258
Soltani, Mehdi H.....	474, 477
Sosnin, P.	64
Suladze, Aleksandra.....	336
Suliman, Mamoun.....	212
Surguladze, Gia.....	308
Svetlana, Eyubova.....	504

T

Tang, Yuhua.....	217
Tashpulatova, N.B.	62
Tekin, Oner	119
Theljani, Foued	376
Topuria, Nino.....	308
Tosun, Suleyman.....	273
Tosun, Umur	12
Treesinthuros, Wasin	304
Tsabadze, Teimuraz	321
Tsamalashvili, Tengiz	321
Turčáni, Milan.....	450
Turkia, Ekaterine.....	308

U

Ulvi Simsek, Mehmet	398
---------------------------	-----

V

Vilums, Sandis	138
Voevodin, I.G.....	62

W

Wijekoon, Janaka.....	207
Wojcik, Waldemar	414

X

Xu, Xinhai.....	217
Xu, Yan.....	99

Y

Yagain, Deepa	159
Yaghmaee Moghaddam, Mohammad Hossein.....	285
Yaghoubi Suraki, Mohsen.....	71
Yaseen, Amer Atta	279
Ye, Yongmao	455
Ye, Zhengmao	455
Yıldırım Okay, Feyza.....	441
Yıldız, Hakan	181

Z

Zacepins, Aleksejs	155
Zautashvili, David.....	333
Zeinolabedini, Zahra	39, 192
Zhang, Hongwu.....	410
Zhu, Tongyu.....	113
Zhukov, S.....	64
Zidi, Salah.....	376