Danco Davcev
Jorge Marx Gómez
Editors

ICT Innovations 2009

Springer

ICT Innovations 2009

Danco Davcev and Jorge Marx Gómez

ICT Innovations 2009

Springer

Prof. Danco Davcev
Ss. Cyril & Methodius University
Fac. Electrical Engineering &
Information Technologies
Karpos 2
1000 Skopje
Macedonia
E-mail: etfdav@feit.ukim.edu.mk

Prof. Dr.-Ing. Jorge Marx Gómez
Universität Oldenburg
Fak. Informatik
Abt. Wirtschaftsinformatik
Ammerländer Heerstr.
114-118
26129 Oldenburg
Germany
E-mail: jorge.marx.gomez@uni-oldenburg.de

ISBN 978-3-642-10780-1 e-ISBN 978-3-642-10781-8

DOI 10.1007/978-3-642-10781-8

Library of Congress Control Number: 2009941532

Typesetting: Data supplied by the authors

Production: Scientific Publishing Services Pvt. Ltd., Chennai, India

Cover Design: WMX Design, Heidelberg, Germany

Printed in acid-free paper

9 8 7 6 5 4 3 2 1

springer.com

Preface

This book is the result of the first International Conference ICT Innovations 2009. The ICT Innovations conference is the primary scientific action of the Macedonian Society on Information and Communication Technologies (ICT-ACT). It promotes the publication of scientific results of the international community related to innovative fundamental and applied research in ICT. Today, ICT has enlarged its horizons and it is practiced under multidisciplinary contexts that introduce new challenges to theoretical and technical approaches.

The ICT Innovations 2009 conference gathered academics, professionals and practitioners reporting their valuable experiences in developing solutions and systems in the industrial and business arena especially innovative commercial implementations, novel applications of technology, and experience in applying recent research advances to practical situations, in any ICT areas. The conference focuses on issues concerning a variety of ICT fields like:

- Multimedia Information Systems
- Artificial Intelligence
- Pervasive and Ubiquitous Computing
- Eco and Bio Informatics
- Internet and Web Applications and Services
- Wireless and Mobile Communications and Services
- Computer Networks, Security and Cryptography
- Distributed Systems, GRID and Cloud Computing

ICT Innovations 2009 Conference was held in Ohrid, Macedonia, in September 28-30, 2009. Local arrangements provided by the members of the Macedonian Society on Information and Communication Technologies – ICT-ACT, mainly consisting of teaching and research staff of Computer Science Department at Faculty of Electrical Engineering and Information Technologies and Institute of Informatics at Faculty of Natural Sciences, both at Ss. Cyril and Methodius University in Skopje, Macedonia.

Editors would like to express their gratitude to Vista group, Pexim Solutions, Seavus, Accent Computers, Genrep Software group, Matrix global, Microsoft Macedonia, Ein-Sof, Faculty of Electrical Engineering and Information Technologies and Faculty of Natural Sciences that sponsored the publication of the present book, and personally to Dr. Slobodan Kalajdziski for his wholehearted support given during the preparation and publication of the book.

Ohrid,
September 2009

Danco Davcev
Jorge Marx Gómez

Organization

ICT Innovations 2009 Organizers

Conference and Program Chair

Dr. Danco Davcev	Ss. Cyril and Methodius University (UKIM), Faculty of Electrical Engineering and Information Technologies, Skopje, Macedonia

Program Committee

Ackovska Nevena	UKIM, Macedonia
Amata Garito Maria	UNINETTUNO - International Telematic University, Italy
Andonovic Ivan	University of Strathclyde, UK
Antovski Ljupco	UKIM, Macedonia
Atanassov Emanouil	IPP BAS, Bulgaria
Bosnacki Dragan	TUE, Netherlands
Cakmakov Dusan	UKIM, Macedonia
Celakovski Sasko	ITgma, Macedonia
Chitkuchev T. Lubomir	Boston University, USA
Dika Zamir	SEEU, Macedonia
Dimitrova Nevenka	Philips Research, USA
Dimov Zoran	Microsoft - Vancouver, Canada
Fullana Pere	ESCI - Barcelona, Spain
Furht Borko	Florida Atlantic University, USA
Gavrilovska Liljana	UKIM, Macedonia
Gievska-Krliu Sonja	GWU, USA
Gligoroski Danilo	Univ. Trondheim, Norway
Grünwald Norbert	Hochschule Wismar, Germany
Gusev Marjan	UKIM, Macedonia
Haak Liane	University Oldenburg, Germany
Hadzi-Velkov Zoran	UKIM, Macedonia
Jonoska Natasha	Univ. of South Florida, USA
Josimovski Saso	UKIM, Macedonia
Junker Horst	IMBC, Germany
Kalajdziski Slobodan	UKIM, Macedonia
Kimovski Goran	SAP, Canada
Kocarev Ljupco	UKIM, Macedonia

Kon-Popovska Margita	UKIM, Macedonia
Kulakov Andrea	UKIM, Macedonia
Kut Alp	9 Eylul University Izmir, Turkey
Lazarova-Molnar Sanja	UAE University, UAE
Margaritis Kostas	Univ. Macedonia, Greece
Markovski Smile	UKIM, Macedonia
Marovic Branko	University of Belgrade, Serbia
Marx Gómez Jorge	Oldenburg University, Germany
Milosavljevic Milan	Singidunum University Belgrade, Serbia
Misev Anastas	UKIM, Macedonia
Mitreski Kosta	UKIM, Macedonia
Mitrevski Pece	UKLO, Macedonia
Mustafa Blerim	MT, Macedonia
Mylonopoulos Yannis	Aristotle University, Greece
Nanevski Aleksandar	Microsoft Research - Cambridge, UK
Olariu Stephan	ODU, USA
Paris Francois	University of Houston, USA
Parychek Peter	Univ. Donau, Austria
Patel Dilip	LSBU, UK
Patel Shushma	LSBU, UK
Profumo Francesco	Politecnico di Torino, Italy
Pudlowski Zenon J.	WIETE - Melbourne, Australia
Radevski Vladimir	NYUS, Macedonia
Sjursen Harold	New York University, USA
Spasovski Mome	UKIM, Macedonia
Stojanov Georgi	American University of Paris, France
Stojcev Mile	Univ. Nis, Serbia
Tasic Jurij	Univ. of Ljubljana, Slovenia
Tochtermann Klaus	Univ. Graz, Austria
Trajanov Dimitar	UKIM, Macedonia
Trajkovic Ljiljana	SFU, Canada
Trajkovik Vladimir	UKIM, Macedonia
Trajkovski Igor	NYUS, Macedonia
Trichet Francky	Nantes University, France
Vasileska Dragica	ASU, USA

Table of Contents

Invited Keynote Papers

Corporate Environmental Management Information Systems – CEMIS 2.0 1
Jorge Marx Gómez

Computational Electronics and 21st Century Education 5
Dragica Vasileska

Wireless Sensor Networks for Cattle Health Monitoring 21
Ivan Andonovic, Craig Michie, Michael Gilroy, Hock Guan Goh, Kae Hsiang Kwong, Konstantinos Sasloglou, and Tsungta Wu

Tools for High Throughput Differential Methylation Study in Cancer . . . 33
Nevenka Dimitrova, Sitharthan Kamalakaran, Angel Janevski, Nilanjana Banerjee, Vinay Varadan, Robert Lucito, and James Hicks

Proceeding Papers

Compensatory Fuzzy Ontology 35
Ariel Racet Valdés, Rafael A. Espin Andrade, and Jorge Marx Gómez

Conceptual Clustering and Analysis of Data from Gynecological Database 45
Cveta Martinovska

System for Prediction of the Winner in a Sports Game 55
Eftim Zdravevski and Andrea Kulakov

Computer Generated News Site – TIME.mk 65
Igor Trajkovski

Applying Bagging Techniques to the SA Tabu Miner Rule Induction Algorithm 75
Ivan Chorbev and Mirjana Andovska

Information/Material Processing Synergy: Flexible Manufacturing and Operating System Metaphor for a Biological Cell 85
Nevena Ackovska and Stevo Bozinovski

Position Estimation of Mobile Robots Using Unsupervised Learning Algorithms 95
Petre Lameski and Andrea Kulakov

Practical Method for Real-Time Path Planning and Optimization for Mobile Robots 105
Saso Koceski and Natasa Koceska

Protein Classification Based on 3D Structures and Fractal Features 115
Georgina Mirceva, Zoran Dimov, Slobodan Kalajdziski, and Danco Davcev

Protein Function Prediction Based on Neighborhood Profiles 125
Kire Trivodaliev, Ivana Cingovska, Slobodan Kalajdziski, and Danco Davcev

Automated Structural Classification of Proteins by Using Decision Trees and Structural Protein Features 135
Slobodan Kalajdziski, Bojan Pepik, Ilinka Ivanovska, Georgina Mirceva, Kire Trivodaliev, and Danco Davcev

Wireless Sensor Networks Localization Methods: Multidimensional Scaling vs. Semidefinite Programming Approach 145
Biljana Stojkoska, Ilinka Ivanoska, and Danco Davcev

On the Complexity of the Greedy Construction of Linear Error-Correcting Codes 157
Dejan Spasov and Marjan Gusev

Vulnerability Assessment of Complex Networks Based on Optimal Flow Measurements under Intentional Node and Edge Attacks 167
Igor Mishkovski, Risto Kojchev, Dimitar Trajanov, and Ljupco Kocarev

Outage Probability of Multi-hop Relay Systems in Various Fading Channels 177
Jovan Stosic and Zoran Hadzi-Velkov

Non-poisson Processes of Email Virus Propagation 187
Miroslav Mirchev and Ljupco Kocarev

Impact of Community Structures on Ad Hoc Networks Performances 197
Sonja Filiposka and Dimitar Trajanov

Rule Induction of Physical-Chemical Water Property from Diatoms Community 207
Andreja Naumoski and Kosta Mitreski

Semantic Interaction in Enterprise Data-Flow Visualization Environments: An Exploratory Study 217
Alberto Morell Pérez, Jorge Marx Gómez, and Carlos Pérez Risquet

A Beehive-Like Multi-agent Solution to Enhance Findability of Semantic Web Services and Facilitate Personalization within a P2P Network 227
Ammar Memari, Mahmoud Amer, and Jorge Marx Gómez

Phonetic Experiment Web Application 237
Anastazija Kirkova-Naskova, Goran Bakraceski, Vladimir Apostolski, and Dimitar Trajanov

Using XAML in Representation of Dental Charts in Electronic Health Record 247
Ivica Marković, Srebrenko Pešić, and Dragan Janković

Enterprise Tomography Driven Governance of Federated ERP in a Cloud 257
Jan Aalmink, Lama Balloul, Jan Glagau, and Jorge Marx Gómez

Composite Index of e-Business Strategy Readiness of the Enterprises in the Republic of Macedonia 265
Marjan Angeleski, Pece Mitrevski, and Margarita Janeska

Performance Evaluation of a New Approach for Automatic Question Production 275
Mile Jovanov and Marjan Gusev

Semantic Supported Modeling and Orchestration of Logistic Integrated Processes, with Focus on Supply Chain: Framework Design 285
Roberto Pérez López de Castro, Dania Pérez Armayor, Jorge Marx Gómez, Inty Sáez Mosquera, and José Antonio Díaz Batista

Web Service Validation within Semantic SOA-Based Model 295
Tariq Mahmoud, Timo von der Dovenmühle, and Jorge Marx Gómez

Auto-production 3D Graphics Content for Mobile Communication 305
Blagica Jovanova, Marius Preda, and Françoise Preteux

On-Line Animation System for Learning and Practice Cued Speech 315
Ivica Arsov, Blagica Jovanova, Marius Preda, and Françoise Preteux

AUM and Enterprise Tomography: New Concepts for Technology Enhanced Learning for Enterprise Systems in Higher Education 327
Liane Haak, Jan Aalmink, and Dirk Peters

Multiplatform Real-Time Rendering of MPEG-4 3D Scenes with Microsoft XNA 337
Sasko Celakovski and Danco Davcev

Fast Classification Scheme for HARDI Data Simplification 345
V. Prčkovska, A. Vilanova, C. Poupon, B.M. ter Haar Romeny, and M. Descoteaux

Experimental Comparison of PSNR and SSIM Metrics for Video Quality Estimation 357
Zoran Kotevski and Pece Mitrevski

Quasigroup String Transformations and Hash Function Design: A Case Study: The NaSHA Hash Function 367
Aleksandra Mileva and Smile Markovski

Performances of Error-Correcting Codes Based on Quasigroups 377
Aleksandra Popovska-Mitrovikj, Smile Markovski, and Verica Bakeva

On the Computational Asymmetry of the S-Boxes Present in Blue Midnight Wish Cryptographic Hash Function 391
Danilo Gligoroski and Vlastimil Klima

Optimization of Adaptive Petri-Net Grid Genetic Algorithm Workflows 401
Boro Jakimovski, Dragan Sahpaski, and Goran Velinov

Massively Parallel Seismic Data Wavelet Processing Using Advanced Grid Workflows 411
Ljupco Jordanovski, Boro Jakimovski, and Anastas Misev

Grids in the Near Future: A Technical and Social Review 421
Margita Kon-Popovska and Anastas Misev

Object Oriented Approach in Computer Aided Process Planning 429
Valentina Gecevska, Franc Cus, and Uros Zuperl

Videoconferencing as Tool of Higher Education 439
Enrica Caporali, Vladimir Trajkovik, and Juna Valdiserri

A New Solution for Workflow and Document Management Used for University Management in WODOMI Project 451
Thomas Biskup, Marjan Gusev, Gjorgji Manceski, and Pece Mitrevski

Author Index 461

Corporate Environmental Management Information Systems – CEMIS 2.0

Jorge Marx Gómez

Department of Business Informatics I, Carl von Ossietzky University Oldenburg,
Ammerländer Heerstr. 114-118, 26129 Oldenburg, Germany
marx-gomez@wi-ol.de

1 Traditional Approach

Dealing with environmental issues in the management level of companies is a relatively new thought that came up within the late eighties. With the upcoming idea on sustainability, viewing ecological, social, and economic issues on the same level, international politics increasingly challenged companies to internalize their impacts on the environment. Furthermore, the development of voluntary eco-management systems like EMAS or ISO 14001 was another fact that shifted the focus more towards companies. Overall, the increasing amount of considerations companies had to put towards environmental issues was initiated externally.

In order to comply with environmental goals, Environmental Management Information Systems (CEMIS), as a special instance of information systems (IS) have been developed. Being introduced mainly to fulfill different external claims, the situation nowadays is the existence of a large number of very specific, heterogeneous solutions in parallel, targeted towards different kind of environmental issues. As such, no true integrative approach exists and even the support of larger-scale problems by CEMIS remains almost exclusively on an operational level.

While there are external factors leading to a benefit in introducing environmental management (e.g. marketing, reduction of resource consumption), it is questionable whether CEMIS focused on an operational level really is feasible to establish sophisticated environmental goals [1]. Additionally, such isolated solutions have all the problems that business informatics tries to prevent with integrated systems, such as data redundancy and spread, heterogeneous user interfaces inefficient communication etc. It is even more surprising that this is a repetition of the mistakes made in the early sixties and seventies of the previous centuries, which nowadays lead to a costly step towards integrating separate, outdated legacy systems.

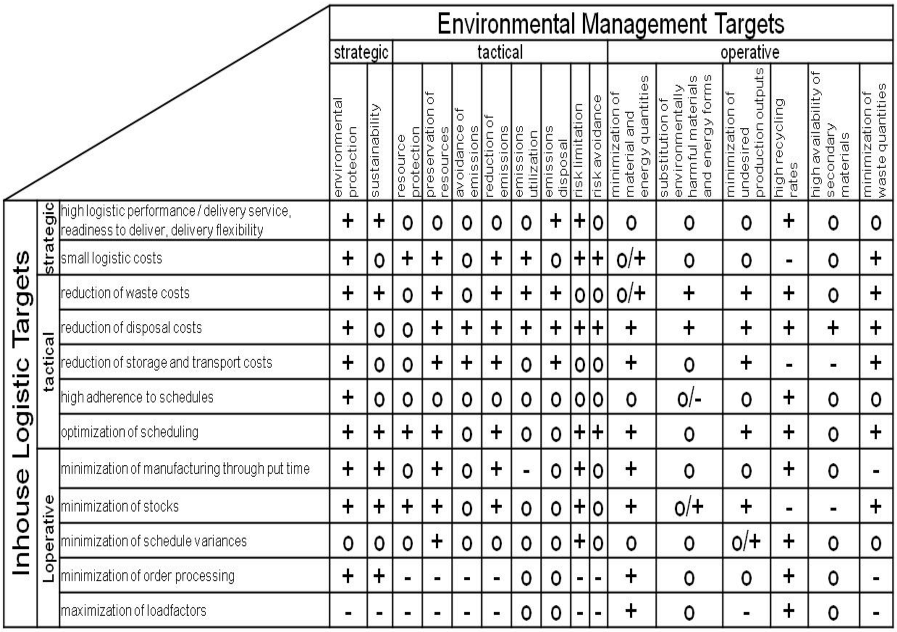
Considering the increasing debate about a sustainable society, there is a growing demand for companies to go beyond strictly adhering to legal requirements. Instead of passively reacting, there is a call to embrace active environmental protection. This includes not only reducing unnecessary waste, but to improve efficiency in production processes overall in terms of production-integrated environmental protection, satisfying the various stakeholder needs. In order to achieve this, a more holistic approach for CEMIS has to be developed in supporting the production chain as a whole.

There are first steps towards integrating corporate environmental protection in the goals of business practice, as there is an increasing insight that emission and waste reduction leads to economic advantages as well. Based on frameworks set by relevant environmental laws, companies can enhance these in order to reduce resource consumption and improve risk assessment. To effectively benefit from environmental protection, environmental management has to become an integral part of the companies' long-term, strategic goals. Establishing only general goals will therefore not suffice, ecological goals have to be specified and made quantifiable where possible in order to have means of evaluation and setting up clear criteria for success, or failure [2].

2 Strategic Environmental Management for In-House Logistics

CEMIS supporting strategic goals are dealing with the integration of ecological and economical goals, increasing the already high complexity of corporate software systems. Their optimization potential is highly related to the system boundaries; the more aspects are considered, the more possibilities are available to improve system optimization in terms of economic and ecological aspects [3]. In the following, the idea for developing a sustainability oriented goal system is proposed. In order to reduce the complexity, the focus will not be laid on an entire company but on the in-house logistics processes. This work is based on the results of C. Lang [4].

In-house logistics are one of the major targets for corporate environmental management, as material and energy flow, as well as transport, handling, and storing processes are handled here. Decision affecting logistics will immediately have effects on the amount and risks of resource usage and emissions. The following figure is setting goals of environmental management and logistics management in contrast, distributed along strategic, tactical, and operative targets:



			Environmental Management Targets															
			strategic		tactical								operative					
			environmental protection	sustainability	resource protection	preservation of resources	avoidance of emissions	reduction of emissions	emissions utilization	emissions disposal	risk limitation	risk avoidance	minimization of material and energy quantities	substitution of environmentally harmful materials and energy forms	minimization of undesired production outputs	high recycling rates	high availability of secondary materials	minimization of waste quantities
Inhouse Logistic Targets	strategic	high logistic performance / delivery service, readiness to deliver, delivery flexibility	+	+	o	o	o	o	o	+	+	o	o	o	o	+	o	o
		small logistic costs	+	o	+	+	o	+	+	o	+	+	o/+	o	o	-	o	+
	tactical	reduction of waste costs	+	+	o	+	o	+	+	+	o	o	o/+	+	+	+	o	+
		reduction of disposal costs	+	o	o	+	+	+	+	+	+	+	+	+	+	+	+	+
		reduction of storage and transport costs	+	o	o	+	+	+	o	+	o	o	+	o	+	-	-	+
		high adherence to schedules	+	o	o	o	o	o	o	o	o	o	o	o/-	o	+	o	o
		optimization of scheduling	+	+	+	+	o	+	o	o	+	+	+	o	+	+	o	+
	Loperative	minimization of manufacturing through put time	+	+	o	+	o	+	-	o	+	o	+	o	o	+	o	-
		minimization of stocks	+	+	+	+	o	+	o	o	+	o	+	o/+	+	-	-	+
		minimization of schedule variances	o	o	o	+	o	o	o	o	+	o	o	o	o/+	+	o	o
		minimization of order processing	+	+	-	-	-	-	o	o	-	-	+	o	o	+	o	-
		maximization of loadfactors	-	-	-	-	-	-	o	o	-	-	+	o	-	+	o	-

Fig. 1. Overall target system of the environmental-orientated in-house logistics.

In cases where the relation between environmental targets and economic targets are in accordance, establishing specific strategies for reaching these goals is relatively easy, as ecological goals can be achieved without further effort when aiming at the economic goals.

Neutral goals are independent and must be enforced and evaluated separately. When goals are conflicting, their benefits have to be judged against each other. In reality, ecological goals will be discarded in favor of economic goals; as such targets are higher ranked within a company's overall strategy. In cases where environmental goals should not be ignored entirely, both targets have to be assessed in comparison to each other in order to find an optimal level of environmental, as well as economic achievement.

The given overview demonstrates that only 12% of the goal relations are in conflict to each other. Compared to the 42% of complementary goals, this is a relatively small number. Yet, for 46% neutral relations there are additional activities necessary.

3 Conclusion

With the evaluation of in-house logistics, an important part of a company has been examined in detail. It has been demonstrated that the realization of a system combining economic and ecologic goals is possible in this area. Although this suggests that such a goal system could be introduced to other parts of a company as well, more research has to be done on this part.

Since the widely distributed idea of corporate environmental protection to be solely imposing costs has been refuted, the future development of CEMIS should highlight this aspect, demonstrating the synergies of economic and ecologic goals. Future research should examine if and how an integrated approach could support both dimensions of sustainability in a single information system.

In terms of the ongoing sustainability debate and the results of this work, the traditional definition of CEMIS has become largely obsolete. Former systems mainly support the idea of operational goals, providing single solutions for different problems with no means of integration. This is true for the majority of CEMIS in companies in use today, showing the same phenomenon that was plaguing information systems in the area of business informatics in the seventies and eighties.

The methods developed with EAI (Enterprise Application Integration) and information management should be applied to the area of CEMIS as well, providing an integrated view on environmental and economic problems in order to overcome the focus on operative management.

Based on the findings mentioned above the following new definitions can be given:

From a strategic point of view, CEMIS 2.0 are such information systems that support the idea of a sustainable development in a company. From a more specific, tactical point of view, CEMIS are such information systems that have a holistic approach towards material and energy efficiency, emission and waste reduction, recycling, stakeholder engagement, and legal compliance.

References

1. Hamschmidt, J.: Wirksamkeit von Umweltmanagementsystemen – Stand der Praxis und Entwicklungsperspektiven, St. Gallen, Diss, p. 59 (2001)
2. Bundesministerium für Umwelt, Naturschutz und Reaktorsicherheit. Umweltbundesamt: Handbuch Umweltcontrolling, München, p. 4 (2001)
3. Heiserich, O.-E.: Logistik. Eine praktische Einführung, 3. Aufl. Wiesbaden, p. 16 (2002)
4. Lang, C.: Ökonomisch-ökologische Zielsysteme als Herausforderung für das Ressourcenmanagement. In: Pietsch, Thomas, Lang, Corinna (Hrsg.) (eds.) Ressourcenmanagement, Berlin, p. 44 (2007)

Computational Electronics and 21st Century Education

Dragica Vasileska

Arizona State University, Tempe, AZ 85287-5706, USA
Tel.: +1 480 965-6651
vasileska@asu.edu

Abstract. The continued scaling of semiconductor devices and the difficulties associated with time and cost of manufacturing these novel device design has been the primary driving force for the significantly increased interest in Computational Electronics which now, in addition to theory and experiments, is being considered as a third important mode in the design and development of novel nanoscale devices. In addition to its significant role in industrial research, modeling and simulation also brings into the picture alternative education modes in which students, by running certain subset of tools that, for example, the nanoHUB offers, can get hands-on experience on the operation of nanoscale devices and can also look into the variation of internal variables that can not be measured experimentally, like the spatial variation of the electron density in the channel in the pre- and post-pichoff regime of operation, electric field profiles which can be used to tailor the electron density to avoid junction breakdown, etc. In summary, Computational Electronics is emerging as a very important field for future device design in both industry and academia.

Keywords: nano-electronics, semiclassical and quantum transport, education.

1 Introduction

In this invited paper we discuss the importance of the Computational Electronics [1] in Today's industry and academia. We first discuss the important modules of Computational Electronics, its application in the design process of future nanoscale devices where we take two examples: the example of investigating self-heating effects in fully-depleted (FD) silicon on insulator (SOI) devices for which we use particle based simulations for the electron transport and energy balance simulator for including the optical and the acoustic bath; and the example of process corner variation in FinFET devices for which we use the Contact Block Reduction (CBR) method to solve the non-equilibrium Green's function (NEGF) formalism in the ballistic limit.

Having explained the importance of Computational Electronics in Industrial applications in Section 2 below, we then turn the attention to the application of the computational electronics to research/academic and purely academic applications (Section 3). As an example of research/academic tool we discuss SCHRED (1D Schrodinger-Poisson solver for MOS capacitors) that is the most cited tool on the nanoHUB (92 citations in scientific research papers). We discuss the user network for this tool and what are the typical applications that SCHRED is used for research. Finally we discuss the SCHRED potential for educational activities.

As examples of purely academic tools installed on the nanoHUB, we discuss the tool-based educational modules ABACUS, AQME and ACUTE, and then we focus on a particularly interesting and important tool from an educational point of view: PCPBT – piece-wise-constant potential barrier tool.

We summarize this paper with highlights on what is the future of Computational Electronics. In that regard we discuss what tools need to be used for modeling future nano-electronic devices in which band-structure effects due to local strain and stresses will necessitate the use of atomistic simulations.

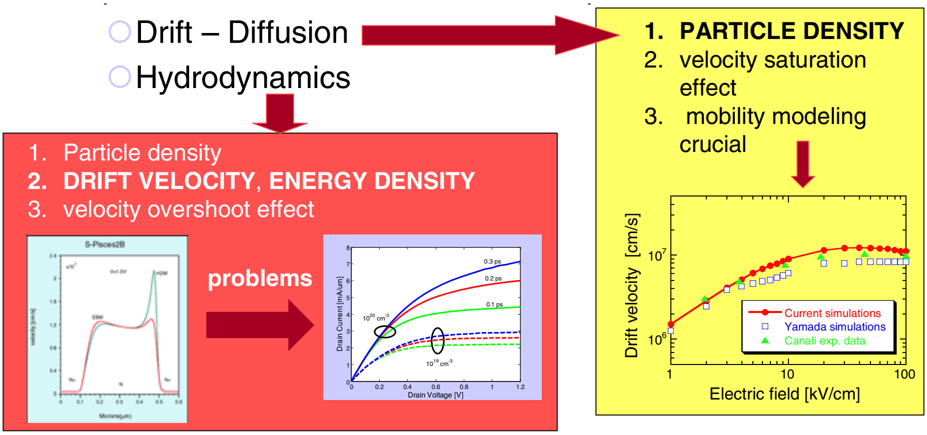
2 What Is Computational Electronics?

Computational Electronics in one sense is integral part of Technology for Computer Aided Design (TCAD) but in other sense is different from TCAD as it involves physically based, not circuit level, device simulation. The physical models chosen can vary in their complexity which, in turn, determines the applicability of the method due to its ability/inability to model certain phenomena. At the semi-classical level, most commonly used methods are drift-diffusion, energy balance and solution of the Boltzmann transport equation using the Monte Carlo method. While drift-diffusion and hydrodynamic models are fluid-type approaches (see Figure 1), the Monte Carlo method is a particle approach. In recent years, due to dominance of non-stationary transport in devices, particle-based approaches are the method of choice. When tunneling and quantum-mechanical size quantization dominate the device behavior, it is preferable to use some of the quantum transport approaches, the most efficient of all being the non-equilibrium Green's functions approach (NEGF) due to Keldysh, Kadanoff and Baym. In the rest of this section we highlight the applicability of particle-based semiclassical transport approaches on the example of self-heating effect in FD SOI devices and afterwards we illustrate the applicability of CBR method for ballistic transport investigations of process parameters variations analysis in FinFET device structures.

2.1 Self-heating Effects in Nano-Scale Devices

The scaling of semiconductor devices into the nanometer regime and the problems associated with further miniaturization of device technologies has resulted into investigation of devices with alternative materials and alternative device designs such as fully-depleted (FD), dual-gate (DG), tri-gate silicon-on-insulator (SOI) and other device designs. The problem with SOI devices is that they exhibit self-heating effects. These self-heating effects arise from the fact that the underlying SiO_2 layer has about 100 times smaller thermal conductivity than bulk Si (1.4 W/m/K). Also, the thickness of the silicon film in nanoscale devices is much smaller than the phonon mean free path which is on the order of 300 nm in bulk silicon. Therefore, boundary scattering becomes dominant scattering mechanism, thus reducing the thermal conductivity value to a fraction of its bulk value. For example, the bulk thermal conductivity in silicon is 148 W/m/K and the thermal conductivity of a silicon film of thickness of

10 nm is 13 W/m/K (a factor of 10 smaller than the bulk value). Also, in thin silicon films the thermal conductivity has smaller temperature dependence because boundary scattering is temperature independent scattering process [2].



○ Drift – Diffusion
○ Hydrodynamics

1. **PARTICLE DENSITY**
2. velocity saturation effect
3. mobility modeling crucial

1. Particle density
2. **DRIFT VELOCITY, ENERGY DENSITY**
3. velocity overshoot effect

problems

Fig. 1. Capabilities of fluid-based approaches. Drift diffusion model, with inclusion of field dependent mobility is able to capture velocity saturation effect. When velocity overshoot becomes important, hydrodynamic model must be used. The problem of the hydrodynamic model is the choice of the proper energy relaxation times.

Also, in analog devices neighboring devices are typically on and if the gate contacts are also biased then there is no heat flow through the gate contact and the side boundaries. In these cases it is appropriate to use Neumann boundary conditions on the side (artificial) boundaries and Neumann boundary conditions on the gate electrode. We consider this the worst case scenario of the operation of the device. Simulation results for the current degradation for different technology of FD SOI devices, summarized in Table 1, are presented in Figure 2 – top panel.

In the case of digital circuits, the devices are rarely on and the use of Dirichlet boundary conditions at the gate and the side boundaries are the appropriate boundary conditions. This corresponds to the best-case scenario of heat removal from the device active region. Simulation results for the current degradation for different technology devices when Dirichlet boundary conditions are applied to the gate and the side boundaries, are shown in Figure 2 – bottom panel.

In order to perform more realistic estimates of the current degradation using temperature and thickness dependent thermal conductivity model we follow the work of Sondheimer [3], that takes into account phonon boundary scattering (by assuming it to be purely diffusive). Namely, the thermal conductivity of a semiconductor film of a thickness a, under the assumption that the z-axis is perpendicular to the plane of the film, the surfaces of the film being at z=0 and z=a, is given by:

$$\kappa(z)=\kappa_0(T)\int_0^{\pi/2}\sin^3\theta\left\{1-\exp\left(-\frac{a}{2\lambda(T)\cos\theta}\right)\cosh\left(\frac{a-2z}{2\lambda(T)\cos\theta}\right)\right\}d\theta \tag{1}$$

where $\lambda(T)$ is the mean free path expressed as $\lambda(T) = \lambda_0(300/T)$ nm where room temperature mean free path of bulk phonons is taken to be $\lambda_0 = 290$ nm. Selberherr [4] has parametrized the temperature dependence of the bulk thermal conductivity in the temperature range between 250K and 1000K. In our case we find that the appropriate expression is (see Figure 3):

$$\kappa_0(T) = \frac{135}{a + bT + cT^2} \quad \text{W/m/K} \tag{2}$$

where a=0.03, b=1.56×10^{-3}, and c=1.65×10^{-6}. Eqs. (1) and (2) give almost perfect fit to the experimental and the theoretical data reported in an Asheghi paper [2] (see Figure 3).

Table 1. Parameters for various simulated device technology nodes (constant field scaling) [1].

L (nm)	tox (nm)	t_{Si} (nm)	t_{box} (nm)	N_{ch} (cm^{-3})	V_{GS}=V_{DS} (V)	I_D (mA/um)
25	2	10	50	1×10^{18}	1.2	1.82
45	2	18	60	1×10^{18}	1.2	1.41
60	2	24	80	1×10^{18}	1.2	1.14
80	2	32	100	1×10^{17}	1.5	1.78
90	2	36	120	1×10^{17}	1.5	1.67
100	2	40	140	1×10^{17}	1.5	1.57
120	3	48	160	1×10^{17}	1.8	1.37
140	3	56	180	1×10^{17}	1.8	1.23
180	3	72	200	1×10^{17}	1.8	1.03

L- Gate Length; tox- Gate Oxide Thickness;

t_{Si}- Active Si Layer Thickness;

t_{box}- BOX Thickness;

N_{ch}- Channel Doping Concentration;

I_D- Isothermal current value (300K).

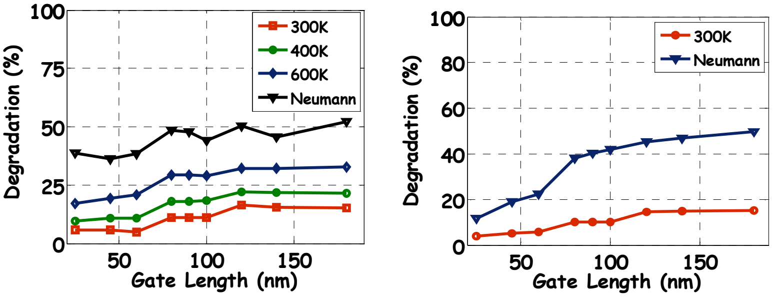


Fig. 2. Left panel - Current degradation vs. technology generation ranging from 25 nm to 180 nm channel length FD SOI devices (Table 1). Isothermal boundary condition of 300K is set on the bottom of the BOX. Parameter is the temperature on the gate electrode. Neumann boundary conditions are applied at the vertical sides. Right panel - Current degradation for the case of Dirichlet boundary conditions at the artificial boundaries. In one case Dirichlet boundary conditions are applied at the gate electrode with T_{Gate}=300 K and in the second case Neumann boundary conditions are applied at the gate electrode.

In Table 2 we compare electro-thermal simulation results for various models for two different device gate lengths (25 nm and 180 nm). Dirichlet boundary conditions are assumed on the gate and back contact (300K), and the other boundaries are treated as Neumann boundary conditions (no heat flow).

Table 2. Absolute values of the currents for: (1) bulk thermal conductivity model, (2) temperature-dependent bulk thermal conductivity model, (3) anisotropic thickness dependent thermal conductivity model, and (4) anisotropic thickness and temperature dependent thermal conductivity model.

Thermal conductivity model	25nm FD SOI (V_{GS}=V_{DS}=1.2V) Current (isothermal):1.824mA/um		180nm FD SOI (V_{GS}=V_{DS}=1.8V) Current (isothermal):1.032mA/um	
	Current (mA/um)	Current Decrease (%)	Current (mA/um)	Current Decrease (%)
142.3 W/m/K	1.714	6.0	0.922	10.7
$\kappa_{bulk}=\kappa_{bulk}(T)$	1.712	6.1	0.915	11.3
anisotropic	1.698	6.9	0.887	14.0
13 W/m/K	1.702	6.7	0.875	15.2

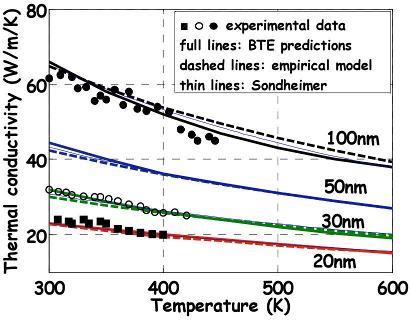


Fig. 3. Silicon film thickness dependence of the average thermal conductivity at T=300 K vs. active silicon layer thickness. Experimental data are taken from the work of Asheghi and co-workers [2].

In Figure 4, we show the temperature maps in the active region of the 25 nm and 180 nm channel length device with the full anisotropic and temperature dependent thermal conductivity model. Compared to earlier results [5], we find that the anisotropic and temperature dependent thermal conductivity model leads to higher lattice temperature profiles at the drain end of the channel and in the channel itself for larger device structures even though the current degradations are very similar. This makes the heat removal process from the drain contact more difficult.

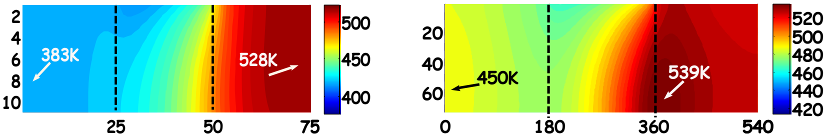


Fig. 4. Lattice temperature for a 25 nm channel length device (left) and a 180 nm gate-length device (right).

2.2 Process Variation in FinFET Devices

Within our group a 2D/3D Contact Block reduction method was developed to investigate ballistic behavior in FinFET device in which gate leakage and quantum-mechanical size quantization effects are important. Details of the methodology used to solve the NEGF formalism are given in Ref. [6].

Figure 5 depicts the channel formation for different fin widths in the on-state (for gate voltage of -0.1V). For a fin width of 12 nm (left panel) simulation results show the presence of distinct channel formed vertically on each side wall of the fin along the gate. Our simulations show that as fin width decreases, the two channels gradually merge into a single channel across the fin. For a fin width of 10 nm two channels are almost merged except for the center region of the fin (Figure 5, middle panel). As fin width shrinks below 10 nm, the fin consists of a single channel resulting from volume inversion (Figure 5, right panel). For even smaller fin width, the channel profile is expected not to change significantly from the volume inversion point of view, but suppression of short channel effects improves up to some cutoff, which will be discussed in subsequent section. In addition, according the scaling rule when shrinking fin width, one has to scale down gate oxide too, in order to optimize the performance, which on the other hand increases gate leakage.

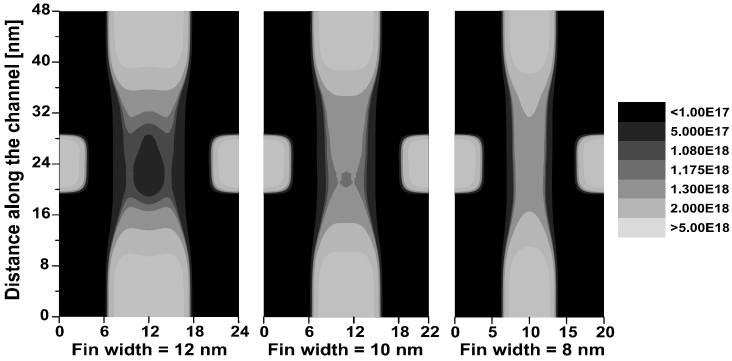


Fig. 5. Electron density for FinFET devices with different fin widths at on-state (V_g=-0.1). The drain and source are at the top and bottom of the plots correspondingly. The gates are on the sides along the horizontal direction of the plots.

The transfer characteristics for different fin thickness ranging from 6 nm to 12 nm are shown in Figure 6. Drain bias of 0.1 V has been used in these simulations. From simulation results it is evident that threshold voltage is negative. However by suitably adjusting the gate work function it is possible to make the threshold voltage positive. We show the transfer characteristics with negative threshold voltage as it resembles the one in experimentally fabricated device in which adjustment of gate work function was not done. It is evident from the simulation results that subthreshold characteristics are greatly affected by the fin width and also narrower fin width results in higher threshold voltage, as compared to the threshold voltage for wider fins.

Figure 7 shows the corresponding variation of subthreshold slope with fin width. With decreasing fin width, the subthreshold slope also decreases, which improves the controllability of short channel effects. After some cutoff value of fin thickness (in this case around 7 nm) subthreshold slope does not decrease that much with decreasing fin width and remains at a nearly constant value. The experimental value of subthreshold slope for 10 nm FinFET is 125 mV/dec. Simulation results for FinFET device with 12 nm fin width shows a subthreshold slope of 120 mV/dec, which is very close to the experimental value. While decreasing fin width to 6-7 nm helps to improve the subthreshold behavior, the corresponding values of the drive-current are reduced. Thus, a compromise must be drawn between these two factors to optimize the device performance. On the other hand, the best subthreshold slope (around 83 mV/dec) we can achieve by just reducing fin width (and keeping the rest of parameters of the experimentally fabricated FinFET constant), is still rather far from the ideal value. For a better performing 10 nm FinFET device, source and drain doping levels should be increased, and oxide thickness should be decreased; the corresponding simulation results with improved subthreshold characteristics will be published separately.

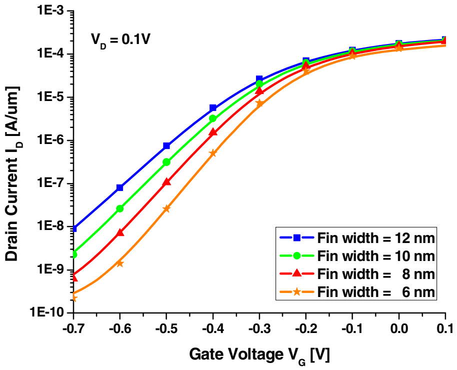


Fig. 6. Variation of transfer characteristics with fin width at V_D = 0.1V

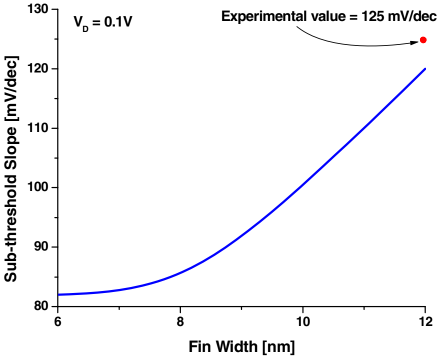


Fig. 7. Subthreshold slope variation with fin width at V_D = 0.1V

To assess the performance degradation of the ultra-scale FinFET device at high temperatures (corresponding to “slow process corner”) we have calculated transfer characteristics with drain voltages of 0.1 and 1.2 V at 400 K. The results of these simulations are shown in Figure 8. For both drain voltages a significant degradation of the subthreshold slope (around 28%) is observed for the considered FinFET device. This result illustrates the high importance of performing slow process corner analysis in nano-scale devices.

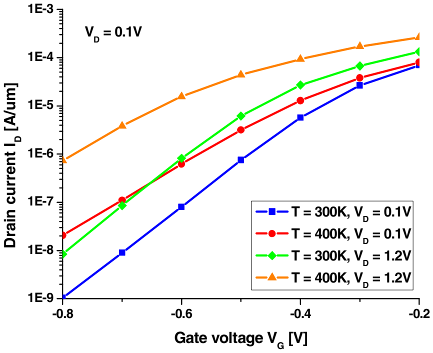


Fig. 8. Simulated FinFET transfer characteristic at V_D=0.1 and 1.2 V, at 300 K (typical process corner) and at 400 K (slow process corner).

3 Computational Electronics and 21st Century Education

As we already discussed in the introduction part of this invited paper, there are several ways in which Computational Electronics is strongly tied with the 21st Century Education: via combined research and educational activities and via purely teaching modules. As an example of a tool that has served both education and research we discuss SCHRED that stands for 1D Schrödinger-Poisson solver.

SCHRED was developed at Arizona State University by Prof. Dragica Vasileska and further extensions to SCHRED to model dual gate devices were achieved as a joint effort between Arizona State University and Purdue University. We want to point out that next generation of SCHRED to model arbitrary materials and multiple conduction and valence bands is currently in development at Arizona State University. This new tool will replace the existing SCHRED tool and will have much more extended capability. Given the visibility that the existing SCHRED tool has at the moment we believe that the next generation SCHRED will be as successful, if not more, than the existing SCHRED tool.

As we already highlighted in the introduction, and we repeat here for completeness, SCHRED has been cited in 92 scientific papers. As such it is THE MOST CITED tool on the nanoHUB. It has been used by users within and outside of the nanoHUB. The demographical usage of SCHRED is depicted in Figure 9.

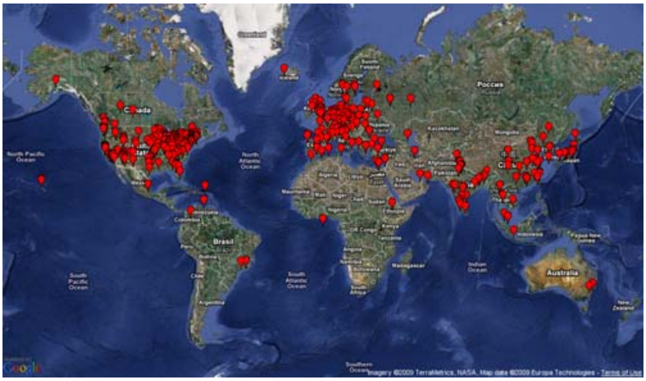


Fig. 9. Demographical Usage of SCHRED.

It is interesting to point out that most users of SCHRED do not necessarily belong to the nanoHUB. We call those outside users. The connection between this group of users is highlighted in Figure 10 below.

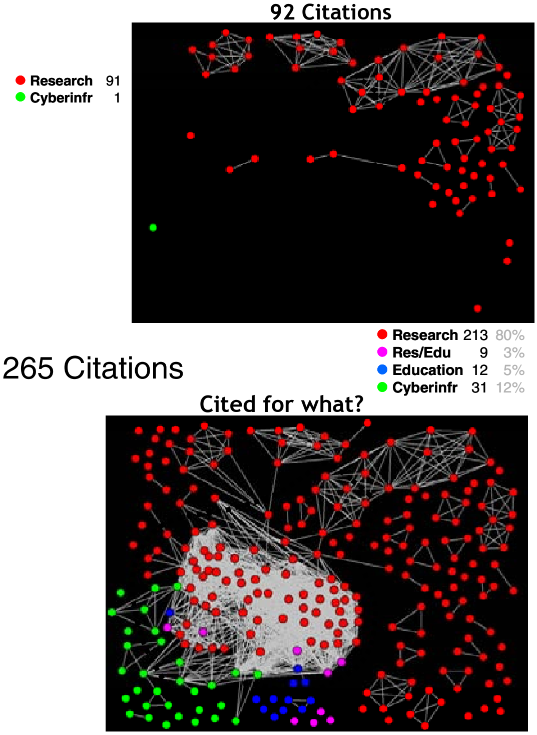


Fig. 10. Top panel – Usage of Schred and connections between various groups. Bottom panel – Usage of all tools installed on the nanoHUB. The heavily connected area depicts the nanoHUB members.

The utility of SCHRED is not in research purposes only. SCHRED is one of the most useful tool for academic purposes as well as it can demonstrate to students studying conventional MOSFET devices (which are integral part of the curricula of every Electrical Engineering school) how with device scaling into the nanometer regime quantum-mechanical size quantization effects become more and more important and why in the smallest devices of research one wants to use metal instead of polysilicon gates due to the considerable poly-gate depletion. Also, SCHRED can be used to estimate the shift in the threshold voltage due to quantum-mechanical size-quantization effects. Results obtained with SCHRED that highlight the poly-gate depletion effect on the total gate capacitance and the shift in the threshold voltage with technology generation are given in Figure 11.

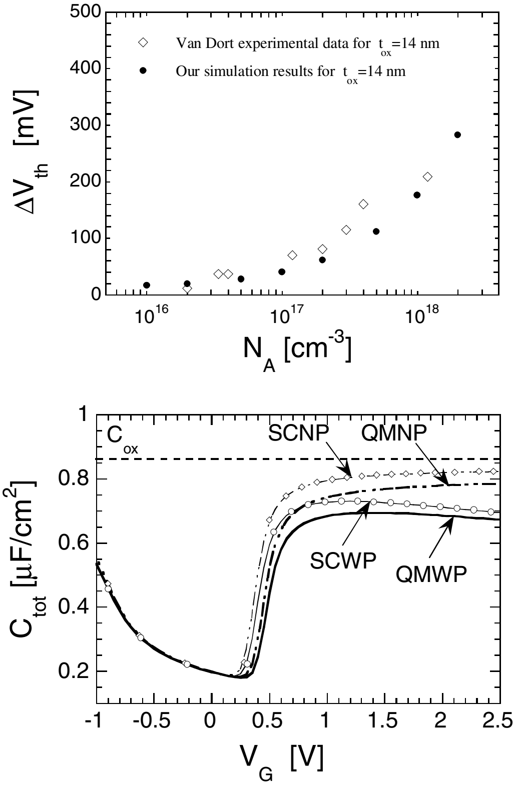


Fig. 11. Top panel - SCHRED simulation data for the shift in the threshold voltage compared to the experimental values provided by van Dort and co-workers. Bottom panel - Variation of low frequency C_{tot} with V_G when using SC and QM description of the charge in the channel, with (WP) and without (NP) the inclusion of the poly-gate depletion. We use N_A=5×10^{17} cm^{-3}, t_{ox}=4 nm and N_D=5×10^{19} cm^{-3}. Also shown here are the semiclassical results obtained with our analytical model (symbols). In the inset we show the variation of C_{poly} with V_G obtained by using the numerical model (solid lines) and our analytical model (symbols).

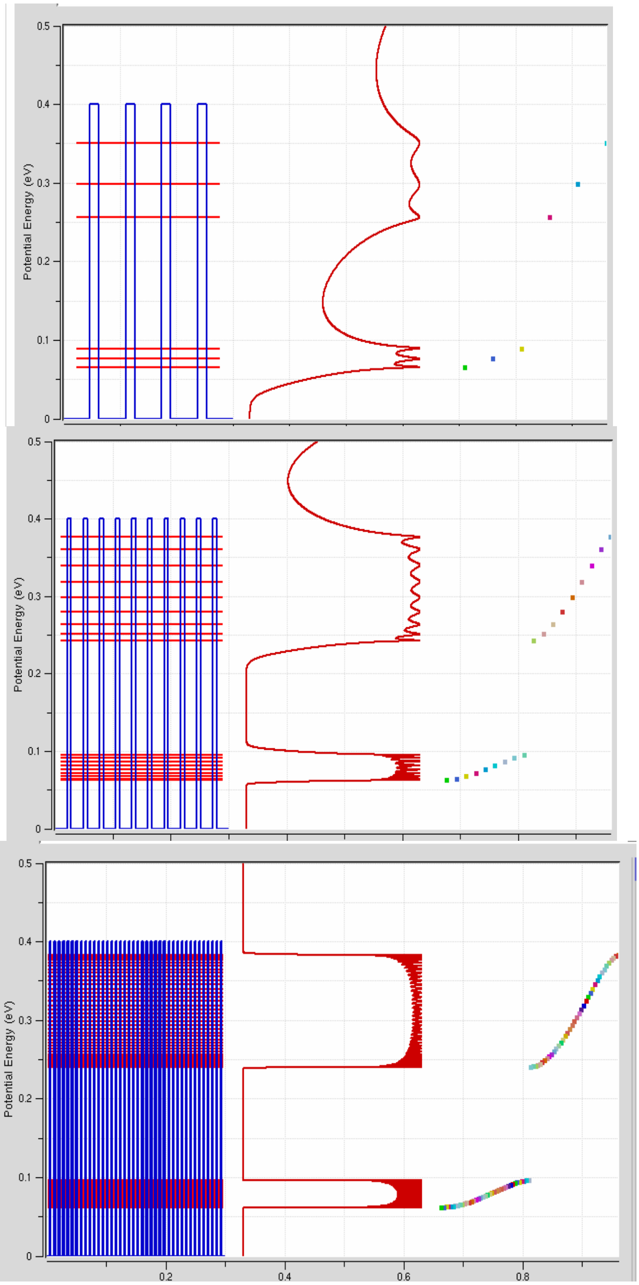


Fig. 12. 4, 10 and 40 potential barriers (top), correspond-ding transmission coefficient vs. energy (middle) and forma-tion of cosine bands (bottom).

Last year the nanoHUB team (Gerhard Klimeck and Dragica Vasileska) promoted tool-based curriculum and as part of this effort ABACUS. AQME and ACUTE were being developed for the purpose to aid conventional classes in semiconductor device theory, quantum mechanics and computational electronics in better understanding the subjects being taught by supplementing them with simulation examples that had to be run by the students taking these classes. The content of each of these composite tools can be found by following this link: www.nanohub.org. In the rest of this section, for illustration purposes we have decided to focus on the capabilities of the PCPBT. Because of limitations of space, in here we describe two examples: (1) the formation of the energy bands and the energy gaps via the example of multiple identical barriers and wells (in the effective mass approximation), and (2) more complex analysis that utilizes tight-binding approach.

In the first example that follows we initially consider four identical barriers and three identical wells. Each well would hold 2 quasi-bound states but because of the interaction between the wells, the degeneracy of the levels is lifted and there are two sets consisting of three quasi-bound states. So, the number of states is preserved and the degeneracy of the levels is lifted. In the next panel we show the situation when there are 10 barriers and 9 wells. Since each well holds 2 resonant states there are total of 2 × 9 = 18 states. As in the previous example, the degeneracy of the states is lifted and we see two bands with nine pronounced resonances which gives exactly 18 states (see Figure 12). On the very right of the second panel we also see that these quasi-bound states start to form cosine bands. The formation of cosine bands is more evident in panel 3 for the case when we have 40 barriers, respectively.

When the barriers are thin and the effective masses have high value, then the effective mass theory fails and one has to consider the complete bandstructure. The following example illustrates this point. We calculate the transmission coefficient of a double barrier structure with barrier height = 0.4 eV, barrier width = 0.5 nm and well width = 4 nm. We also consider the situation of a Si/SiGe material system for which we can assume m*=m$_0$. We make two simulation runs: one using effective mass theory and transfer matrix approach and the second one using tight-binding method and transfer matrix approach. The results of these simulations are shown in Figure 13.

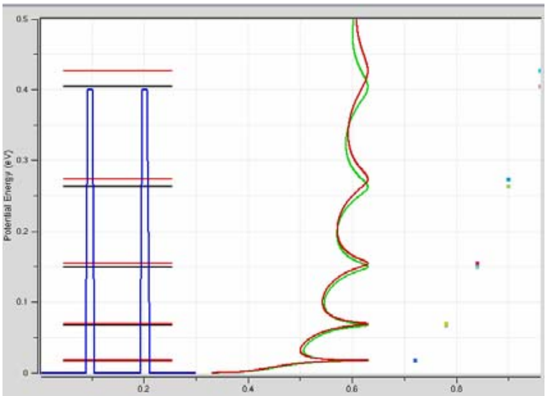


Fig. 13. Example of a symmetric double barrier structure with barrier height = 0.4 eV, barriers widths = 0.5 nm and well width = 4 nm. We use effective mass vs. tight-binding theory.

From the results presented in Figure 13, where we focus on the position of the resonances and the transmission coefficient data it is clear that the tight binding effective mass is higher which lowers the resonances. In another example when the effective mass was much smaller, such is the case of a GaAs/AlGaAs structure, the effective mass and tight binding results were much closer to each other. This example clearly illustrates that for large effective mass, in order to find the proper resonances in a structure, one has to use the tight binding approach.

4 The Future of Computational Electronics

In this invited paper we gave a brief description of currently most important and most physically based semiclassical and quantum transport approaches. It is important to note that because of these developments, device simulation has achieved significantly higher maturity level than process simulation. In fact, particle-based device simulators can capture the essential physics up to ballistic transport regime and, when quantum interference effects start to dominate device behavior, quantum transport simulators based on either direct solution of the Schrödinger equation or its counterpart, the Green's functions, have been developed which, with the recent progress of state of the art computers, can simulate 3D nanoscale devices within a reasonable time-frame.

However, nanoelectronic device simulation of the future *must* ultimately include both, the sophisticated physics oriented electronic structure calculations and the engineering oriented transport simulations. Extensive scientific arguments have recently ensued regarding transport theory, basis representation, and practical implementation of a simulator capable of describing a realistic device.

Starting from the field of molecular chemistry, Mujica, Kemp, Roitberg, Ratner [7] applied tight-binding based approaches to the modeling of transport in molecular wires. Later, Derosa and Seminario [8] modeled molecular charge transport using density functional theory and Green's functions. Further significant advances in the understanding of the electronic structure in technologically relevant devices were recently achieved through *ab initio* simulation of MOS devices by Demkov and Sankey [9]. Ballistic transport through a thin dielectric barrier was evaluated using standard Green function techniques [10,11] without scattering mechanisms. However, quantum mechanical simulations of electron transport through 3D confined structures, such as quantum dots, have not yet reached the maturity (it is important, for example, for simulating operation of the next generation quantum dot photodetectors). Early efforts of understanding the operation of coupled quantum dot structures were rate equation based [12,13,14] where a simplified electronic structure was assumed.

Whereas traditional semiconductor device simulators are insufficiently equipped to describe quantum effects at atomic dimensions, most *ab initio* methods from condensed matter physics are still computationally too demanding for application to practical devices, even as small as quantum dots. A number of intermediary methods have therefore been developed in recent years. The methods can be divided into two major theory categories: atomistic and non-atomistic. Atomistic approaches attempt to work directly with the electronic wave function of each individual atom. *Ab initio* methods overcome the shortcomings of the effective mass approximation; however, additional approximations must be introduced to reduce computational costs. One of the critical questions is the choice of a basis set for the representation of the electronic wave function. Many approaches have been considered, ranging from traditional numerical methods, such as finite difference and finite elements, as well as plane wave expansions [15,16,17], to methods that exploit the natural properties of chemical bonding in condensed matter. Among these latter approaches, local orbital methods are particularly attractive. While the method of using atomic orbitals as a basis set has a long history in solid state physics, new basis sets with compact support have recently been developed [18,19], and, together with specific energy minimization schemes, these new basis sets result in computational costs which increase linearly with the number of atoms in the system without much accuracy degradation [20,21]. However, even with such methods, only a few thousand atoms can be described with present day computational resources.

NEMO3D uses an empirical tight-binding method [22,23] that is conceptually related to the local orbital method and combines the advantages of an atomic level description with the intrinsic accuracy of empirical methods. It has already demonstrated considerable success [24,25] in quantum mechanical modeling of electron transport as well as the electronic structure modeling of small quantum dots [26]. NEMO3D typically uses sp3s* or sp3d5s* model that consists of five or ten spin degenerate basis states, respectively. Note that for the modeling of quantum dots, three main methods have been used in recent years: $k{\cdot}p$ [27,28], pseudopotentials [19], and empirical tight-binding [20].

As already discussed in Section 3, there are a number of methods developed by solid state theorists over the last several decades to address the issue of quantum transport in nano-devices. Among the most commonly used in nanostructure calculations schemes are the Wigner-function approach [29], the Pauli master equation [30], and the non-equilibrium Green's functions (NEGF) [31,32]. The growing popularity of the latest (sometimes referred to as the Keldysh or the Kadanoff–Baym) formalism is conditioned by its sound conceptual basis for the development of the new class of quantum transport simulators [33]. Among its doubtless advantages are the clear physical conceptions, rigorous definitions, well-developed mathematical apparatus and flexibility of the algorithmization.

In summary, from the discussion above it follows that *the ultimate goal of semiconductor transport calculation of future nanoscale devices will be to* ***merge*** *the* ***3D quantum transport approaches*** *with* ***ab-initio band structure calculations***. This will ensure the most accurate simulation and better understanding carrier transport and operation of novel nano-device structures.

Acknowledgements

This work was performed under the support of two NSF grants: ECCS-0901251 and ECCS-0701926.

References

1. Vasileska, D., Goodnick, S.M.: Computational Electronics. Morgan and Claypool, San Francisco (2006)
2. Liu, W., Asheghi, M.: J. Appl. Phys. 98, 123523–1 (2005)

3. Sondheimer, E.H.: Advances in Physics 1(1) (January 1952); reprinted in Advances in Physics 50, 499–537 (2001)
4. Palankovski, V., Selberherr, S.: Journal Microsystem Technologies 7, 183–187 (November 2001)
5. Raleva, K., Vasileska, D., Goodnick, S.M., Nedjalkov, M.: IEEE Transactions on Electron Devices 55(6), 1306–1316 (June 2008)
6. Khan, H.R., Mamaluy, D., Vasileska, D.: IEEE Trans. Electron Devices 54(4), 784–796 (2007)
7. Mujica, K., Roitberg, R.: J. of Chem. Physics 104, 72–96 (1996)
8. Derosa, S.: J. of Phys. Chemistry B 105, 471 (2001)
9. Demkov, A., Sankey, O.: Phys. Rev. Lett. 83, 2038 (1999)
10. Demkov, A., Zhang, L., Loechelt, G.: J. of Vac. Sci. and Techn. B 18, 2388 (2000)
11. Demkov, A., Zhang, Drabold: Phys. Rev. B 6412, 5306 (2001)
12. Klimeck, G., Lake, R., Datta, S., Bryant: Phys. Rev. B 50, 5484 (1994)
13. Klimeck, G., Chen, Datta, S.: Phys. Rev. B 50, 2316 (1994)
14. Chen, et al.: Phys. Rev. B 50, 8035 (1994)
15. Canning, A., Wang, L.W., Williamson, A., Zunger, A.: J. of Comp. Physics 160, 29 (2000)
16. Wang, L.W., Kim, J.N., Zunger, A.: Phys. Rev. B 59, 5678 (1999)
17. Williamson, A.J., Wang, L.W., Zunger, A.: Phys. Rev. B 62, 12963 (2000)
18. Martin, R.: Phys. Rev. B 1, 4005 (1970)
19. Sankey, O., Niklewski, D.J.: Phys. Rev. B 40, 3979 (1989)
20. Ordejón, P., Drabold, D.A., Grumbach, M.P., Martin, R.M.: Phys. Rev. B 48, 14646 (1993)
21. Ordejón, F., Galli, G., Car, R.: Phys. Rev. B 47, 9973 (1993)
22. Vogl, P., Hjalmarson, H.P., Dow, J.D.: J. Phys. Chem. Solids 44, 365 (1983)
23. Jancu, J.M., Scholz, R., Beltram, F., Bassani, F.: Phys. Rev. B 57, 6493 (1998)
24. Bowen, R.C.: IEDM 1997, p. 869. IEEE, New York (1997)
25. Klimeck, G., et al.: VLSI Design 8, 79 (1997)
26. Lee, J., Klimeck, G.: Phys. Rev. B 63, 195318 (2001)
27. Pryor: Phys. Rev. B 57, 7190 (1998)
28. Stier, Grundmann, Bimberg: Phys. Rev. B. 59, 5688 (1999)
29. Brodone, P., Pascoli, M., Brunetti, R., Bertoni, A., Jacoboni, C.: Phys. Rev. B 59, 3060 (1998)
30. Fischetti, M.V.: Phys. Rev. B 59, 4901 (1998)
31. Haque, A., Khondker, A.N.: J. Appl. Phys. 87, 2553 (2000)
32. Guan, D., Ravaioli, U., Giannetta, R.W., Hannan, M., Adesida, I., Melloch, M.R.: Phys. Rev. B 67, 205328 (2003)
33. Datta, S.: Superlattices and Microstructures 28, 253 (2000)

Wireless Sensor Networks for Cattle Health Monitoring

Ivan Andonovic, Craig Michie, Michael Gilroy, Hock Guan Goh,
Kae Hsiang Kwong, Konstantinos Sasloglou, and Tsungta Wu

Centre for Intelligent Dynamic Communications,
Department of Electronic and Electrical Engineering, University of Strathclyde,
204 George Street, Glasgow, United Kingdom

Abstract. This paper investigates an adaptation of Wireless Sensor Networks (WSNs) to cattle health monitoring. The proposed solution facilitates the requirement for continuously assessing the condition of individual animals, aggregating and reporting this data to the farm manager. There are several existing approaches to achieving animal monitoring, ranging from using a store and forward mechanism to employing GSM-based techniques; these approaches only provide sporadic information and introduce a considerable cost in staffing and physical hardware. The core of this solution overcomes the aforementioned drawbacks by using alternative cheap, low power consumption sensor nodes capable of providing real-time communication at a reasonable hardware cost. In this paper, both the hardware and software have been designed to provide real-time data from dairy cattle whilst conforming to the limitations associated with WSNs implementations.

1 Introduction

The farming industry contributes essential revenue to many economies throughout the world e.g. in the UK in year 2006, total value was £14.5 billion and the sector uses around three quarters of the country's land and employs over half a million people. [1]. The two indelible incidents in 1996 and 2001 caused by Bovine Spongiform Encephalopathy (BSE) and Foot and Mouth Disease (FMD) respectively were estimated to have cost the UK economy £13billion in total [2]. Animal health and condition monitoring is now becoming evermore crucial to the wider farming industry as both known and new diseases pose a risk of the global spread of diseases. It is thus important to develop reliable monitoring systems that report a range of animal health conditions back to the farmer or stockman in a timely manner. Current animal monitoring systems only allow data to be downloaded once at a fixed point e.g. during the milking process where sensor nodes attached to dairy cows are downloaded to the data sink through a receiver installed within the milking parlour. This kind of system has an obvious fundamental restriction i.e. the data can only be downloaded at a fixed location. However, only a subset of animals follow the ordered pattern of dairy cattle rendering this approach inapplicable to other animals such as beef cattle, sheep and horses which in turn bounds the market segments the solution can address. This system also exhibits a relatively long response time (delay) as the detected event can only be reported or observed at pre-designated time intervals *viz.* milking time. This potentially will cause unacceptable delays in terms of reacting to critical events like the onset of disease and some events such as oestrus behaviour will not be reported within the appropriate timescales.

This paper therefore reports on the development of a new generation of animal monitoring systems that allow sensory data to reach a farm control system in a timely fashion. Such systems can be considered for both indoor and outdoor farm environments and is applicable a range of animals i.e. cow, sheep, pig, horse, etc. The contributions of this paper are three fold comprising: a prototype designed with due consideration to cost, hardware and software complexity; (2) two novel mobility support schemes proposed for cattle monitoring applications in particular; an evaluation of the performance through field experiments to demonstrate the robustness and effectiveness of the system.

2 Related Activities and Challenges

Various researchers [3, 4, 5] have been using wireless networks in the form of neck mounted sensory collars to track animals' activities and monitor their health conditions with varying degrees of success. One noteworthy system is the 'ZebraNet' [3]. The devices mounted on the zebra transfer all measurements (GPS position) to all other devices within range. A user could then download historical position data from multiple animals by approaching a single zebra. In another example [4], the authors deploy a set of static and mobile sensors. The static nodes measure properties such as soil moisture while mobile nodes carried by livestock study animal behaviours; similar to ZebraNet they use a *store and forward* approach. Such an approach is not scalable due to limited storage space on the device; furthermore the maintenance costs are high since the retrieval of the aggregated measurements requires direct human intervention. Another approach to retrieve data from animal mounted devices utilised an existing GSM infrastructure which facilitates real-time communication [5]. However, battery life concerns aside, this approach becomes prohibitively expensive when monitoring large numbers of animals i.e. the typical cost of a collar is approximately 1700Euros.

To design an appropriate communications platform for the animal monitoring requires an approach beyond the scope of conventional wireless network concepts. As the system design space is constrained by the following factors:

1. Farming environment: open grazing farm fields or ranges are typically used by the farmer to keep animals in a wide open area where they roam freely. The size of a farm can be up to sq kilometres typically in UK and hundreds sq kilometres in Australia and US. Due to size of an open grazing farm, it is almost impossible to have a communication system that provides complete end-to-end coverage.
2. Animal movement: a wireless reception point only provides a data download facility within its radio effective zone *viz.* the area covered by its antenna reception. This radio effective zone typically ranges from 10metres to 100meters diameter. The animals may wander in and out this effective zone freely hence it presents an *ad hoc* series of windows of opportunity for the sensor telemetry

system to successfully transmit a data packet. A communication platform not only needs to support this ad hoc data downloading but also needs to have the capability to download sensory data captured on an animal that may have fallen sick and therefore may be outside the reception zone.

3. Collars are the fundamental front-end devices in many animal monitoring systems, measuring the condition of the animal by for example, sampling the ambient temperature, movement and/or a range of bio-markers. Although the size of a collar and complexity of the resultant electronics packaging varies according to the functionality, in general the collar has to be small in size, light weight, and low cost to be routinely (and affordably) applied. No standards exist but a reasonable maximum for an ear tag for cattle will be of the order of 100g. The weight limit for a collar based device is more likely to be constrained by the pressure and irritation caused by the collar and a reasonable limit for permanent mounting should not exceed 1kg.
4. Radio interference caused by animals: cattle are generally fed in herds, which increases massively the interference surface area, which in turn seriously affects radio performance as a consequence of signal absorption by animals [6]. The hardware and protocol designs must thus take into consideration this interference issue.

3 Animal Behaviour

Domestic cattle are essentially descended from prey species from which they inherit herding traits that benefit them both socially and from a general welfare perspective. However, herds are not uniformly spaced and do not always move as a single collective. Individual animals have different social standings within herds which will influence the animals they are most likely to associate with. As a consequence, the herd may break up into independent sub-herds. This raises an additional question for a WSN; how rapidly network topologies are likely to change and the possibilities of the nodes moving out of range of each other and of base stations. In order to anticipate the extent and rate of such changes, the behaviour of the herd needs to be captured and modelled. In previous studies this has been attempted using collar mounted GPS transponders, one example [7], used GPS to assess the behaviours of 14 free ranging Zebu cows in western Niger. Samples of position were taken at 0.1Hz so that displacement and rate of position change could be used to determine grazing coverage.

With this work in mind, two 24 hour periods of GPS fixes taken at 3 minute intervals from a herd of 14 Limousin and Angus cross beef cows free-ranging on a farm in Midlothian, Scotland, equipped with collar mounted transponders were recorded [8]. Days from summer 2006 were selected at random from a set where the herd had good satellite coverage and there were no major interventions from farm staff. This data set is used to answer two questions that relate to the viability of the WSN: firstly, what is the range between the animals and a base station and secondly, what are the distances between animals most likely to be.

To answer these, it is better to examine the way the quantities of interest are distributed rather than to use point estimate statistics since they may be skewed or multimodal. The best way of producing an accurate picture of a probability density function is to use a kernel density estimator such as a Parzen Window [9]. In the kernel density estimates shown in Figure 1, the probability distribution of the distance of herd from base station on 4th August is shown in Figure 1(a) and distance of herd from base station on 8th August is shown in Figure 1(b). On the 4th of August, the most likely distance from the base-station is around 50metres with a second, smaller mode at around 90metres. On the 8th August this is entirely concentrated around the 80m mark.

With low power wireless sensor networks, range from the base station can be an issue even in fields of the size of those on UK farms. In this trial, although animals have been found to stray to ranges of up to 300metres from a base station, the most likely distances they were to be found at were between 50metres-90metres [8]. The above density functions show the minimum inter-cow distance on 4th August in Figure 1(c) with the minimum inter-cow distance from the 8th August given in Figure 1(d). In both cases the majority of observations are below 40metres with the maximally likely distance being at around 10metres [8].

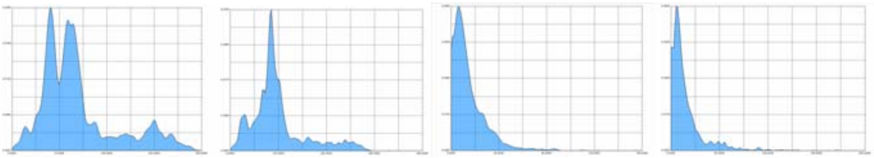


Fig. 1. (a) probability distribution of the distance of herd from base station 4th August; (b) probability distribution of the distance of herd from base station 8th August; (c) minimum inter-cow distance 8th August; (d) minimum inter-cow distance 8th August.

4 Hardware Design and Realization

Wireless sensor nodes are known for their constrained capacities in terms of energy, limited computational power and low memory capability e.g. a MICAz node [10] is powered by two alkaline AA batteries and has one 4MHz processor with 128kB of memory and 4kB of RAM. Given these inherent limited capacities, the implementation of a cattle monitoring solution raises specific and severe challenges. In the following, various factors related to hardware design, such as radio frequency selection, radio propagation behaviour on animal’s body, battery lifetime will be investigated so that a solution which can obtain real-time data from diary cattle whilst conforming to the limitations associated with WSNs implementations can be developed.

4.1 Frequency Selection

In the livestock monitoring system, the animal is free to roam and wireless technology is considered the only feasible method to establishing and maintaining connectivity between base stations and collars attached to the cattle. Radio link quality between transmitter and receiver plays a major role in the performance of any radio network. The radio connectivity range is determined by frequency, transmitted power, antenna characteristics and the radio propagation channel. The main significant factor is the

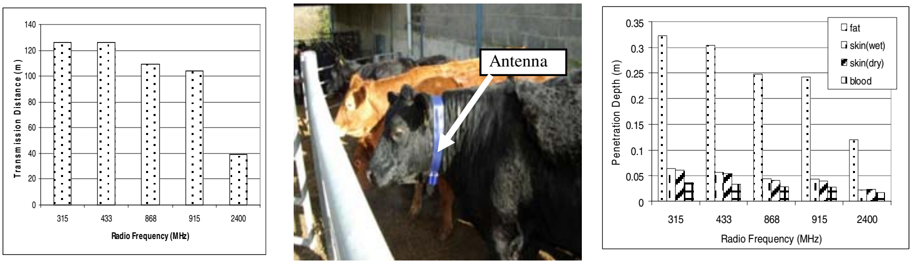


Fig. 2. (a) Transmission range for different frequencies; (b) Antenna placement; (c) Penetration depth for different frequencies.

selection of radio frequency. Figure 2(a) shows achievable transmission range for different frequencies with the power transmission set at 0dBm for a received signal strength at -65dBm. The antenna gains for the collar and the base station are 1dBi and 6dBi respectively. The collar antenna height is 1m above ground and the base station antenna height is 4metres.

4.2 Penetration Depth

The most common collar formats are either a neck collar or a leg collar [11]. Figure 2(b) shows a common antenna placement on a collar (on the side) and the resultant impact that the cow has on the transmission capability. Essentially, the collar can only successfully transmit a packet if the antenna is facing to the base station since the signal can not penetrate the animal’s body. The locations of the collar represent a compromise. Locations on the side of the neck would minimise the effect of shadowing by the animal wearing the collar. However, it also would be susceptible to shadowing from other neighbouring animals in the immediate vicinity. An estimate of signal penetration through an animal can be made using the electrical properties of body tissues [6, 12]. The properties of mammalian tissue are expected to be similar between species. Figure 2(c) summarises the penetration depth at five ISM-band frequencies. Penetration at 2.4GHz is less than 2.5cm in fleshy tissues. Although deeper penetration occurs at 315MHz, the width of a cow’s neck is approximately 0.25metres.

4.3 Bandwidth versus Battery Lifetime

Battery life is a limited resource for collar use, and in the system design it is always useful to target as long an operational life span as possible. For cattle monitoring, a collar may be expected to operate up to at least five years without battery replacement. Commonly, battery conservation is achieved by firstly reducing power consumption owing to radio transmission and secondly through implementing an appropriately low duty cycle [13]. On top of these factors, this paper also presents an evaluation of the impact of radio capacity on battery lifetime. The radio component is switched into low power mode as soon as radio transmission for data download is complete, the length of the transmission cycle being directly dictated by radio channel capacity. Table 1 provides a battery lifetime comparison between a low capacity

(MICA2, sub 1GHz, 76.8kbps) and a high capacity (MICAz, 2.4GHz, 250kbps) sensor nodes. In this study, the sensor nodes are required to transmit a total amount of 100kbytes of data to a base station in a daily basis; the radio channel is turned off as soon as the transmission is finished.

Table 1. Battery lifetime comparison for low and high radio capacity sensor nodes.

	Data load 100kB		Data load 1MB	
	MICA2	MICAz	MICA2	MICAz
delay (s)	10.41666667	3.2	104.1666667	32
current transmit (mA)	10.4	17.4	10.4	17.4
current stand by (mA)	0.0002	0.00002	0.0002	0.00002
radio cycle				
current transmit	4.513888889	2.32	45.13888889	23.2
current stand by	0.719913194	0.071997	0.719131944	0.071973333
total	5.233802083	2.391997	45.85802083	23.27197333
battery capacity	lifetime (months)			
1000 mAh	6.368856293	13.93536	0.726881203	1.432338068
2000 mAh	12.73771259	27.87071	1.453762405	2.864676137
5000 mAh	31.84428146	69.67678	3.634406013	7.161690342

4.4 Prototype System

Our prototyping system comprises collars with antenna diversity (Figure 3(a)) and base stations (Figure 3(b)). The system can be extended to increase network coverage easily with a solar powered relay router as shown in Figure 3(c); this device is also equipped with two antennas. The first antenna is an omni-directional antenna mainly used for receiving packets from the surrounding area, whereas the second antenna is a ceramic patch directional antenna [14] capable of beaming a received packet with a narrower angle over a longer range. A number of relay routers can be used to form a backhaul radio link allowing data captured from surrounding animals to be relayed back to a central location.

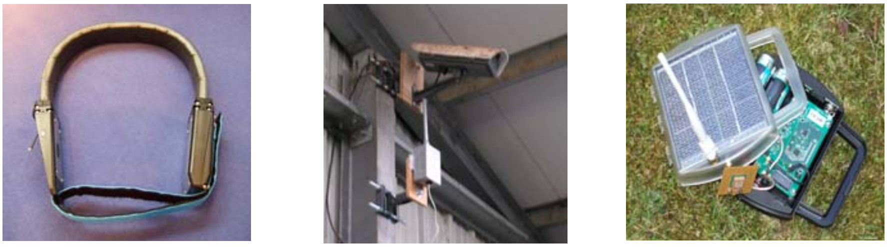


Fig. 3. (a) A collar with antenna diversity; (b) An installed base station at the farm; (c) Solar powered router with antenna diversity.

4.5 Antenna Height

Radio link quality between transmitter and receiver plays a major role in the performance of any radio network. The radio connectivity range is determined by frequency, transmitted power, antenna characteristics and the radio propagation channel. A single, line-of-sight, path between transmitter and receiver seldom exists in a real world environment. In open environments, received signal strength may be very sensitive to the strength of the ground reflected propagation path. In the case of a strong ground reflection receive, and transmit, antenna heights (above ground level) will have a large impact on received signal strength depending on whether interference between direct and reflected paths is constructive or destructive. A simple two-path model can be used to describe (and predict) this effect [15]. Using this model the received signal power, P_r, at distance d is given by:

$$P_r(d) = \frac{P_t G_t G_r h_t^2 h_r^2}{d^4 L} \qquad (1)[15]$$

where P_t is transmitted power, G_t and G_r are transmit and receive antenna gains (as ratios, not in dBi), h_t and h_r are the heights of transmit and receive antennas above the ground, d is the distance between transmit and receive antennas and L accounts for any losses not represented by the two-path model. P_t, h_r, G_t and G_r are determined by selection and configuration of the communications hardware whereas h_t varies according to animal size. On a farm the transmit antenna height is approximately 1.2metres (for a standing animal of average height). In principal the height for the base station (h_r) antenna could be optimised (for the expected value range of d providing this range is not too large) to ensure close to constructive interference between direct and ground reflected signals. This optimum height is given by:

$$h_r = n\frac{\lambda d}{4h_t} \qquad (2)[15]$$

where λ is wavelength and n is an odd integer which would normally be chosen to 1. Choosing a value of n > 1 would mean a taller and therefore more expensive antenna tower [15]. This is unlikely to be advantageous unless, for example, greater antenna elevation resulted in a significantly reduced probability of shadowing.

Measurements of received power have been made and compared power predicted by Equation 1 and Equation 2. Assuming that the animals are roaming in an area within 40metres of the base station, the transmit antenna height was set at 1.2metres, consistent with the collar being worn by an animal of average size. Figure 4(a) shows a comparison of theoretical values (two-ray model) with experimental results of received signal power at the receiver. In Figure 4 the height of receiver's antenna is varied from 0.25metres to 2.25meters and the best received signal strength is obtained at receiver's antenna height of 1meter. The terrain, on which the experiments were conducted, was on grass covered open ground (approximately 5cm tall). 95 % confidence intervals for mean received power have been calculated using 300 RSSI samples. The prediction assumes terrain permittivity and conductivity values of 6 and 0.1 respectively, typical of grassland used for grazing cattle. More information regarding the impact of ground surface on signal propagation and reflection can be referred from Fresnel reflection equation [16]. Figure 4(b) shows the (theoretical) base station antenna height for three (optimum values of n).

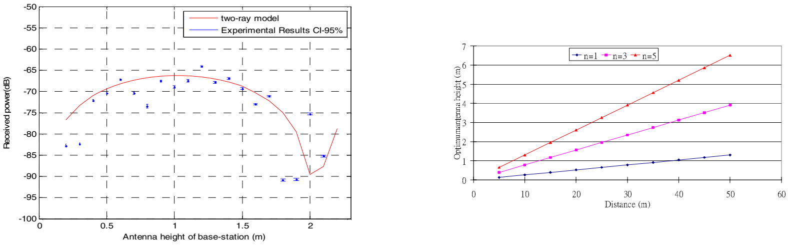


Fig. 4. (a) Optimum antenna height at base station; (b) optimum antenna height (*n*=1, 2, 5).

5 Data Collection

The connectivity between each collar is often sporadic leading to unstable routing paths which in turn result in increased packet delay. To lessen the impact of node mobility, an Implicit Routing Protocol (IRP) is designed particularly for cattle monitoring systems. The proposed IRP operates according to the following two phases: the configuration phase and the data forwarding phase. During the configuration phase, the Base Station (BS) periodically floods a TIER message throughout the entire network. This TIER message contains a BS's ID field, and a hop count field. The hop count field is used to track the number of hops the TIER message has traversed from the base station, the TIERS being numbered starting from the base station. A collar in a given tier, *n,* represents the n^{th} tier away from the BS. This critical information is the *TIER ID.* As animals are free to move, the BS is required to send TIER messages periodically at intervals of T_s to maintain the correct configuration. At the data forwarding phase, if the collar is required to report measured data back to the BS, it will form a packet containing its current *TIER ID* and measurement data. This packet is then broadcasted. Only receiving collars with lower *TIER IDs* are required to respond with an acknowledgment (ACK) packet. These collars, after acknowledging the source collar, will broadcast the packet. Conversely, receiving collars with equal or higher *TIER IDs* will discard the received data immediately. This forwarding rule will then repeat until the data arrives at the BS. Thus the measured data will move one hop closer to the BS at each forwarding stage.

The IRP performance was further investigated through experiments. The IRP was implemented on the MICAz node using TinyOS (TinyOS, 2009); the test-bed was a 3-hop network with one source node, one base station (BS) and *N* pairs of intermediate relay nodes. Figure 5 depicts the average packet delay performance and Figure 6 the received packet rate of test bed configuration *N* = 4, and in each tier there are 4 relay nodes. During each experiment, the source node generates 10,000 packets at an interval of 250ms, each packet containing 85bytes in the payload.

In order to simulate movement, an asynchronous random "on/off" mechanism was implemented. A sensor node in "off" mode represents a cow movement out of communication range; and when a sensor node is switched to "on" mode, it represented a cow entering the communication range. This "on/off mechanism" was characterised by an "off" probability P_{off} which determined the probability the sensor node remains in "off" mode. Figure 5 and Figure 6 both show that network performance is severely

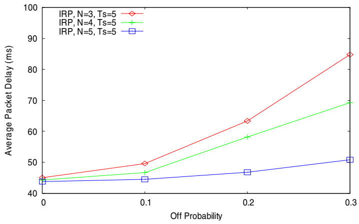


Fig. 5. Average Packet Delay.

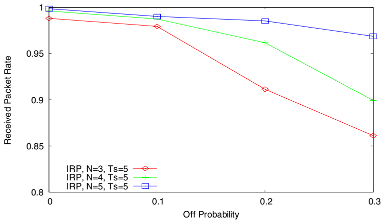


Fig. 6. Received Packet Rate.

impacted as P_{off} increases. However, performance is improved when the number of sensor nodes in each tier increases.

The data collector, - also referred to as “data mule” – is a mobile/portable device that can be brought in to the field where the cattle gather for the purpose of collecting data [17]. This scheme is useful when the IPR cannot establish a path. This could happen if there are big gaps in distance between two groups of animals. The hardware structure of a data collector is essentially identical to a normal sensor node but with additional memory and power capacities. These additional add-ons allow the data collector to constantly scan for any new sensor node which comes into contact, and downloads stored data from that sensor node. In the farm environment, the collector can be carried by a well-trained dog or mounted on to a tracker and sent out into the fields. After data has been collected from the field, the data collector can be connected to a computer directly for data downloading.

The communication protocol of the data collector can be divided into two parts: discovery process and the data transfer process. During discovery, the data collector determines if there are any animals in the vicinity by periodically broadcasting a beacon. This beacon will be acknowledged by any sensor node that receives it. When the collector receives a response back from a sensor node it will send out an acknowledgement and the discovery process will be terminated at this point. The data transfer process will then be initialised, in which data is exchanged between sensor node and data collector. Each of the packets sent by the sensor node will be acknowledge by the collector, hence by receiving an acknowledgement from the collector, the sensor node can remove this stored data from its memory and be secure that this data has been successfully transmitted. Figure 7 shows the protocol flowchart.

The performance of data collector can be improved by reducing the duty cycle of the sensor node. By reducing the sleep period, the sensor node is more likely to “hear”

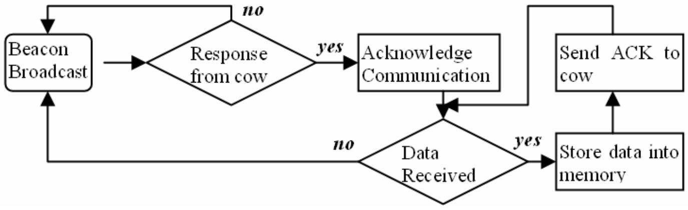


Fig. 7. Data collector flow chart diagram.

the beacon sent out by the collector, so data can be transmitted before the collector moves out of range. Of course, by reducing the duty cycle the sensor node will increase its power consumption level. Figure 8 shows the collector in action.

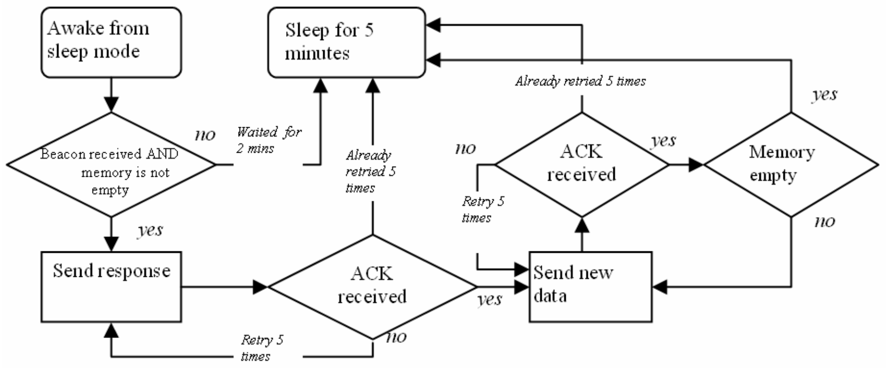


Fig. 8. Sensor node operational protocol.

6 Conclusions

A real-time design for a concept cattle health monitoring systems using wireless sensor networks has been presented. Challenges imposed from adaptation of wireless sensor networks in agriculture and farming have been studied and evaluated. The main difficulty of adapting wireless sensor network into cattle monitoring lies in the capability supporting each node's mobility as a consequence of animal movements. A detail analysis of herd distribution based on 14 Limousin and Aberdeen Angus cross beef cows in a working farm is provided. With the knowledge of the herd mobility, two tailored networking schemes are proposed facilitating real-time data download. These schemes enable up-to-date animal condition states to be fed back to the farm manager, with the goal of improving animal health/welfare and operational efficiency via more informed decision-making.

Acknowledgements

The authors acknowledge the support from ITI Scotland and the UK Engineering and Physical Sciences Research Council for parts of this work.

References

[1] Defra. Defra, UK - Department of Environment, Food, and Rural Affairs (2009), `http://www.defra.gov.uk/` (accessed January 28, 2009)

[2] Mathews, K., Buzby, J.: Dissecting the Challenges of Mad Cow and Foot-and-Mouth Disease. Agricultural Outlook, 4–6 (2001)

[3] Zhang, P., Sadler, C.M., Lyon, S.A., Martonosi, M.: Hardware Design Experiences in ZebraNet. In: Proc. of the 2nd International Conference on Embedded Networked Sensor Systems, Baltimore, MD, USA (2004)

[4] Sikka, P., Corke, P., Valencia, P., Crossman, C., Swain, D., Bishop-Hurley, G.: Wireless Adhoc Sensor and Actuator Networks on the Farm. In: Proc. of the 5th ACM International Conference on Information Processing in Sensor Networks, Nashville, Tennessee, USA (2006)

[5] Mayer, K., Ellis, K., Taylor, K.: Cattle Health Monitoring Using Wireless Sensor Networks. In: Proc. of the 2nd IASTED International Conference on Communication and Computer Networks, Cambridge, Massachusetts, USA (2004)

[6] Gabriel, S., Lau, R.W., Gabriel, C.: The dielectric properties of biological tissues: III. Parametric Models for the Dielectric Spectrum of Tissues. Physics in Medicine and Biology 41, 2271–2293 (1996)

[7] Schlecht, E., Hülsebusch, C., Mahler, F., Becker, K.: The Use of Differentially Corrected Global Positioning System to Monitor Activities of Cattle at Pasture. Applied Animal Behaviour Science 25 (2004)

[8] Kwong, K.H., Goh, H.G., Michie, C., Andonovic, I., Stephen, B., Mottram, T., Ross, D.: Wireless Sensor Networks for Beef and Dairy Herd Management. In: The 2008 American Society of Agricultural and Biological Engineers Annual International Meeting (ASABE AIM), Providence, Rhode Island, USA (2008)

[9] Schwager, M., Anderson, D.M., Butler, Z., Rusa, D.: Robust Classification of Animal Tracking Data. Computers and Electronics in Agriculture 56 (2007)

[10] Crossbow. MICAz 2.4GHz, Crossbow Technology (2009), `http://www.xbow.com/` (accessed January 28, 2009)

[11] Westfalia. GEA Farm Technologies (2009), `http://www.westfalia.com/uk/en/` (accessed January 28, 2009)

[12] IFAC. Dielectric Properties of Body Tissues: HTML clients (2009), `http://niremf.ifac.cnr.it/tissprop/htmlclie/htmlclie.htm` (accessed January 28, 2009)

[13] Ye, W., Heidemann, J.: Medium Access Control with Coordinated Adaptive Sleep-ing for Wireless Sensor Networks. IEEE/ACM Transactions on Networking 12(3), 493–506 (2004)

[14] TDK. TDK Corporation - Americas (2009), `http://www.tdk.com` (accessed January 28, 2009)

[15] Glover, I., Grant, P.: Digital Communications. Prentice-Hall, Englewood Cliffs (1998)

[16] Rappaport, T.S.: Wireless Communications Principles and Practices, 2nd edn. Prentice Hall, Englewood Cliffs (2002)

[17] Anastasi, G., Conti, M., Monaldi, E., Passarella, A.: An Adaptive Data-transfer Protocol for Sensor Networks with Data Mules. In: Proc. of the IEEE International Symposium on a World of Wireless, Mobile, and Multimedia Networks, Espoo, Finland (2007)

Tools for High Throughput Differential Methylation Study in Cancer

Nevenka Dimitrova[1], Sitharthan Kamalakaran[1], Angel Janevski[1§],
Nilanjana Banerjee[1], Vinay Varadan[1], Robert Lucito[2], and James Hicks[2]

[1] Philips Research North America, 345 Scarborough Road, Briarcliff Manor, NY 10510, USA
[2] Cold Spring Harbor Laboratory, 1 Bungtown rd, Cold Spring Harbor, NY 11724, USA

1 Background

Advancement in molecular bioinformatics research is generating an overwhelming amount of data in various modalities. In particular, breast cancer research is advancing at a great pace with the latest transcriptomic, genomic and epigenomic studies. A new approach of exploring differential DNA methylation correlated to cancer reveals great new tools for looking at the DNA information to predict functional downstream regulatory effects on tumor suppressor genes and oncogenes.

2 Methods

We discuss the basic computational problems involved in genomic studies. We have undertaken a high throughput high-resolution microarray study that reveals the methylation status of over 25,000 CpG islands in the human genome. We used this array method to examine the information from breast tumors for potential alterations in the methylation status of CpG islands, both those associated with gene promoters, as well as those non-promoter associated islands.

We developed a platform for molecular signature discovery and clinical decision support that relies on genomic and epigenomic measurement modalities as well as clinical parameters such as histopathological results and survival information. Our **P**hysician **A**ccessible **P**reclinical **A**naly**y**tics **A**pplication (PAPAyA) integrates a powerful set of statistical and machine learning tools that leverage the connections among the different modalities. It is easily extendable and reconfigurable to support integration of existing research methods and tools into powerful data analysis and interpretation pipelines.

3 Results

Using data from high throughput DNA methylation profiling we have identified differentially methylated genomic loci which are methylated in cancer cell lines and at the same time their respective gene expression is not above background. We also compare these candidate loci with differential methylation in tumor vs. normal samples. It is well known that methylation of CpG islands associated with gene promoter regions can affect the expression of the proximal gene, and methylation of non associated CpG islands correlates to genomic instability. Interestingly, there are different methyl binding proteins that are associated with these loci. In addition, we evaluate the regulatory potential and certain regulatory elements such as transcription factors associated with the differentially methylated loci.

4 Conclusion

Our software platform, PAPAyA, enables analysis of data from various data modalities in high throughput molecular studies such as methylation and gene expression studies. In addition, we enable the formulation of new clinical hypotheses which help in decision support of elucidating molecular profiles for therapy decisions.

Compensatory Fuzzy Ontology

Ariel Racet Valdés[1], Rafael A. Espin Andrade[1], and Jorge Marx Gómez[2]

[1] Management Studies Center, José Antonio Echeverría University
Calle 114 No. 11901, Marianao. Ciudad de La Habana, Cuba
{aracet,espin}@ind.cujae.edu.cu
[2] Dpt. Business Informatics I, Carl von Ossietzky University Oldenburg
Ammerländer Heerstr. 114-118, 26129 Oldenburg, Germany
marx-gomez@wi-ol.de

Abstract. Nowadays, to have relevant information is an important factor that contributes favorably to the decision making process. The usage of ontologies to improve the effectiveness in obtaining information has received special attention from researchers in recent years. However, the conceptual formalism supported by ontologies is not enough to represent the ambiguous information that is commonly founded in many domains of knowledge. An alternative is to incorporate the concepts of compensatory fuzzy logic in order to handle the uncertainty in the data, which take advantage of the benefits it provides for the formal representation of uncertainty. We present in this paper the formal definition of "Compensatory Fuzzy Ontologies" and attempt to bring to light the need for enhanced knowledge representation systems, using the advantages of this approach, which would increase the effectiveness of using knowledge in the field of decision making.

Keywords: Ontologies, Compensatory Fuzzy Logic, Decision Making Process, Compensatory Fuzzy Ontologies.

1 Introduction

Information is one of the strategic resources of any organization. To have a minimum of outstanding information could be an important factor to contribute favorably to the decision making process that takes place along the entire value chain of a company, as well on the training of the people involved in that process.

Making changes without information implies ignorance on the problem dimension, impossibility to monitor the advances and inability to evaluate the results [1]. Not all the information is vital for the decision making process, for this process is important to have the tangible and vague information relative to the environmental factors, organizational factors, decision – specific factors and the decision process characteristics. [2]

Human cognitive limitations make very difficult, questionable, subject to inconsistencies, to manage the multiple dimensions of the conflict when the number of attributes increases beyond a few [3]. This will influence significantly the decision making process, where the decision-maker must assess and manage several issues relevant to the decision.

It is necessary to have tools that are able to assist managers in the decision making process. The usage of ontologies to improve the effectiveness in obtaining information has received special attention from researchers in recent years. However, the conceptual formalism supported by ontologies is not enough to represent the ambiguous information that is commonly founded in many domains of knowledge [4], where decision making is not an exemption: while more strategic is the decision, greater is the amount of uncertainty and vagueness the decision maker has to deal with. It is well understood that relations among real life entities are always a matter of degree, and are, therefore, best modeled using fuzzy relations [5].

The incorporation of the concepts of compensatory fuzzy logic inside ontologies in order to take advantage of the use of vagueness in the knowledge domain, allows the enhancement of the formal representation and its employment in knowledge management.

2 Ontology and Fuzzy Ontology

Ontology is a conceptualization of a domain into a human understandable, machine readable format consisting of entities, attributes, relationships and axioms [6]. It is used as a standard knowledge representation for the Semantic Web.

There are several fuzzy concepts that we cannot conceive using conventional ontologies [7]. To handle uncertainty of information and knowledge, one possible solution is to incorporate fuzzy theory into ontology. Then we can generate fuzzy ontologies, which contain fuzzy concepts and fuzzy memberships. The fuzzy ontologies are capable of dealing with fuzzy knowledge [8].

A fuzzy ontology Fo consists of four elements $(C;A^C;R;X)$, where C represents a set of concepts, A^C represents a collection of attributes sets, one for each concept, and $R =(R_T;R_N)$ represents a set of relationships, which consists of two elements: R_N is a set of non-taxonomy relationships and R_T is a set of taxonomy relationships. X is a set of axioms. Each axiom in X is a constraint on the concepts and relationships attribute values or a constraint on the relationships between concept objects. [4]

Some new approach of multivalued fuzzy logic call compensatory fuzzy logic can improve some limitations that classical point of view of norm and co-norm has. Next sections introduce some of these criticisms and solutions, relevant to enhance the fuzzy ontologies role in Knowledge Management and Decision Making.

3 Fuzzy Logic and Modeling of Decision

One way to implement the "principle of gradualness" - essential property of Fuzzy Logic - is the definition of logics where the predicates are functions of the universe X in the interval [0,1] and the conjunction, disjunction, negation and involvement operation are defined in such a way that when restricted to the domain {0,1} we obtain the Boolean Logic. The different ways of defining the operations and their properties determine different multivalued logics that are part of the Fuzzy Logic Paradigm [9].

The use of a set of different operators with properties that generalize the bivalued logic would seem to be the natural way to model decision problems from the language. In fact, applications in the field of decision making has been made basically from the operator concept, rather than multivalued logic [10]. However, this way to address the decisions does not provide the best base to exploit the capacity of Fuzzy Logic for knowledge transformation and decision maker preferences in logical formulas.

There are two main features that hinder the use of logic-based approaches in decision modeling:

- The associative property of conjunction and disjunction operators used
- No compensation among truth values of basic predicates when the compound predicates veracity is calculated using the operators.

The associativity property of a large part of the operators used for aggregation determines that objectives hierarchy trees, which represent different preferences, produce the same truth values of its compound predicates. Under the associativity property both trees in Fig. 1 represent the same preferences, something inappropriate in a decision making model. It is obvious, for example, that the target x has greater relevance in the tree on the right than in left.

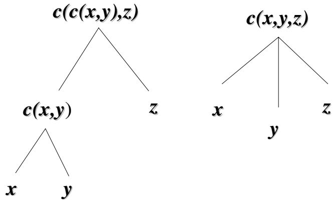


Fig. 1. Objectives hierarchy trees

The lack of compensation is an obstacle for a model that seeks to norm or to describe the reality of decision making; the classic approaches of decision theory, the base of normative thinking, includes models such as the additive ones, which accept the compensation without limits. Descriptive approaches accept partial compensation, which seems more akin to the reasoning of the actual agents. In this way the total lack of compensation and associativity are important limitations of operators frequently used for the addition of preferences.

The above suggests that is desirable the creation of non-associative multivalued logic systems, that facilitate the truth values compensation between basic predicates. Compensatory fuzzy Logic is a multivalued logic that meets these requirements. This is proposed as a decision logical approach, which joins decision modeling and reasoning.

4 Compensatory Fuzzy Logic

Let x = (x1, x2 ,..., xn) be any element of the Cartesian product $[0,1]^n$. A quartet of continuous operators (c, d, o, n), c and d $[0,1]^n$ in [0,1], or $[0,1]^2$ in [0,1] and n of [0,1] in [0,1] represents a compensatory logic, if it satisfies the following set of axioms:

Compensation Axiom, Symmetry or Commutativity Axiom, Strict Growth Axiom, Veto Axiom, Fuzzy Reciprocity Axiom, Fuzzy Transitivity Axiom and De Morgan's Laws.

Operators c and d are called conjunction and disjunction, respectively. Operator o is called fuzzy-strict ordering, and n the negation operator.

Among the multiple aggregation operators found in the literature the only operator which satisfies the axioms of compensation, symmetry, strict growth and veto is the geometric mean:

$$c(x_1, x_2, \ldots., x_n) = (x_1 . x_2 \ldots x_n)^{1/n} . \tag{1}$$

and consequently with the De Morgan's laws, correspondent disjunction will be:

$$d(x_1, x_2, \ldots., x_n) = 1 - ((1-x_1)(1-x_2)\ldots.(1-x_n))^{1/n} . \tag{2}$$

From the above it follows that the quartet of operators formed by the geometric mean and its dual as conjunctive and disjunctive operators, together with the order (3) and negation n(x) = 1− x represent a compensatory logic.

The implication can be defined in general as (4) or (5).

$$o(x, y) = 0.5[C(x) - C(y)] + 0.5 \tag{3}$$

$$i_1(x, y) = d(n(x), y) . \tag{4}$$

$$i_2(x, y) = d(n(x), c(x, y)) . \tag{5}$$

thereby generalizes the truth tables of Boolean logic in two different ways.

Equivalence is defined from operator *i* as

$$e(x, y) = i(x, y) \wedge i(y, x) . \tag{6}$$

The universal and existential quantifiers must be introduced naturally from the conjunction and disjunction operators, for which introducing already the selected operators, we have:

$$\underset{x \in U}{\forall} p(x) = \underset{x \in U}{\wedge} p(x) = \sqrt[n]{\underset{x \in U}{\prod} p(x)} = \\ = \begin{cases} \exp\left(\dfrac{1}{n}\displaystyle\sum_{x \in U} \ln\big(p(x)\big)\right) & if\ x\, p(x) \neq 0 \\ 0 & \text{in any other case} \end{cases} \tag{7}$$

$$\underset{x \in U}{\exists} p(x) = \underset{x \in U}{\vee} p(x) = 1 - \sqrt[n]{\underset{x \in U}{\prod(1 - p(x))}} = \\ = \begin{cases} 1 - \exp\left(\dfrac{1}{n}\displaystyle\sum_{x \in U} \ln\big(1 - p(x)\big)\right) & if\ x\, p(x) \neq 0 \\ 0 & \text{in any other case} \end{cases} \tag{8}$$

For the case of limited sets over R^n, universal and existential quantifiers are defined naturally from the concepts of conjunction and disjunction respectively, moving on to the continuous case through integral calculus [12]:

$$\forall x\, p(x) = \begin{cases} e^{\frac{\int_X \ln(p(x))\, dx}{\int_X dx}} & if\ p(x) > 0 \text{ for any } x \in X \\ 0 & \text{in any other case} \end{cases} \tag{9}$$

$$\exists x\, p(x) = \begin{cases} 1 - e^{\frac{\int_X \ln(1-p(x))\, dx}{\int_X dx}} & if\ p(x) > 0 \text{ for any } x \in X \\ 0 & \text{in any other case} \end{cases} \tag{10}$$

4.1 Relation between Compensatory Fuzzy Logic and Boolean Logic

The formulas for the Propositional Calculus of Compensatory Logic (PCCL) are functions of operators c, d, n and i. Consistent with the definition given by expression (9), any function (11) from the PCCL is considered valid if f (x)> 0 for any element and (12).

$$.f{:}[0,1]^n \rightarrow [0,1]\,. \tag{11}$$

$$e^{\frac{\int_{[0,1]^n} \ln(1-p(x))\, dx}{\int_{[0,1]^n} dx}} > \frac{1}{2}\,. \tag{12}$$

According to the predicate calculus introduced through the definitions of the quantifiers, it's satisfied the following theorem of Compatibility with Boolean Logic:

Theorem 1: The valid formulas of the PCCL are exactly the ones from Boolean Propositional Calculus (BPC) for any of the two selections of the implication operator (i1 o i2), correspondent to formulas (4) and (5) respectively.

4.2 Compound Inference

The next theorem is the key for a new type of inference which join together the logical and statistical inference:

Theorem 2: Suppose M denote a random sample of the universe U. If p(x)>0 for all x in U, then the universal proposition (13) is normally distributed as (14) where σ^2 is the variance of ln(p(x)) in U and (15) is the mean of ln(p(x)) over U.

$$\underset{x \in M}{\forall}\, p(x) = \exp\left(\frac{1}{n}\sum_{x \in M} \ln(p(x))\right) \tag{13}$$

$$N(u, \frac{\sigma^2}{n}) \tag{14}$$

$$u = \frac{1}{n}\sum_{x \in U} \ln(p(x)) \tag{15}$$

This result is obtained from definition (7) using the central limit theorem, and allows the estimation of the truth of a universal proposition using a sample. It can be used to reason using a compound inference, which translates from the language for modeling scenarios using predicates of the CFL, and estimates the truth about the universe using the veracity of these predicate on the sample. This inference can also be employed using the same scheme of reasoning, applying Monte Carlo techniques to estimate the accuracy of a universal proposition.

5 Compensatory Fuzzy Ontology

A Compensatory Fuzzy Ontology is a conceptualization of a domain into a human understandable, machine-readable format consisting of fuzzy concepts and non-fuzzy concepts, fuzzy properties and non-fuzzy properties, fuzzy relationships and non-fuzzy relationships, axioms, instances, using compensatory fuzzy logic to obtain the truth values of fuzzy elements expressed through fuzzy predicates.

The concepts, properties and relationships keep the exact same definitions and play the same roles as in a classical ontology. The concepts, properties and fuzzy relations are defined by compensatory fuzzy logic, used to represent elements of the fuzzy area modeled.

- A fuzzy concept $\tilde{C}$ is defined as a fuzzy set whose membership function over the universe U is the associated property c that is defined through a predicate c(x), $x \in U$; the correspondent crisp set is C=$\{x \in U: c(x) > 0.5\}$ that is a non fuzzy class.
- Axioms are compound predicates using defined concepts.

The definition of Compensatory Fuzzy Ontology allows taking into account Theorem 1 to ensure that the results of a bivalued reasoning, from the consideration that the classes are true when their value is greater than 0.5, are the same as if the compensatory fuzzy operators are applied to the calculation of the veracity of predicates and consequently the membership degree to a class. In this second case for each instance we get the truth values of each of the classes.

The knowledge representations obtained using Compensatory Fuzzy Ontology can be used for querying and knowledge discovering [3][11][13]; this means that it is possible from data and other properties to automate the process of ontology enrichment.

The fulfillment of Theorem 2 allows the use of the Compensatory Fuzzy Ontology to estimate real values of the universal proportions from samples, the compound inference associated with the theorem makes easy to work with a sample of instances to infer knowledge.

The knowledge represented trough fuzzy ontologies should be exploited by a reasoner using the definitions of all the elements in CFL. The reasoner can answer some requests trough simple evaluation of predicates and discover knowledge by the use of searching methods.

6 Case Study: Ontology for Competitive Enterprise

The following ontology represents the basic domain knowledge for competitive enterprise. The represented knowledge by the ontology is a consensual knowledge of BIOMUNDI consulting firm.

Following are presented the statements and their translation into the language of predicate calculus:

A firm is competitive in a product line at a given market if 1) the economy of the company is solid and 2) its technology is advanced and 3) it is very strong in the product line at the market

1. A company is financially sound if it has a good financial state and good sales. If the financial state is a not so good, it must be offset by very strong sales.
2. A company has an advanced technological position if your current technology is good and also owns patents, or it has products in research and development, or significant amounts of money devoted to this activity. If their technology is somewhat behind, then it must have many patents, many products in research and development, or spend very substantial amounts of resources to this effort.
3. A company is strong in a product line, if you have strength in the market, has a diversified product line and it is independent of the supplier.

The following notation associates concepts to names of predicates that will be defined using the "translation" of the statement.

- *c(x)*: Enterprise x is competitive
- *s(x)*: Enterprise x has solid economy
- *t(x)*: Enterprise x has advanced technological position
- *l(x)*: Enterprise x is strong in the product line
- *f(x)*: Enterprise x has good financial state
- *v(x)*: Enterprise x has good sales
- *g(x)*: Enterprise x has a good technology nowadays
- *p(x)*: Enterprise x is the owner of patents
- *i(x)*: Enterprise x many products in research and development
- d(x): Enterprise x or spend very substantial amounts of resources in research and development
- *m(x)*: Enterprise x has strength in the market
- *vl(x)*: Enterprise x has a diversified product line
- *ip(x)*: Enterprise x is independent of the supplier

The compound predicate c(x) obtained from the expressed definition of Competitive Enterprise is the property that defines the fuzzy class $\tilde{C}$ that represents competitive enterprise:

$$c(x) = s(x) \wedge t(x) \wedge l^2(x) \tag{16}$$

where the use of the exponent 2 means "very" and is used as a modifier like is usual in Fuzzy Logic literature. The correspondent no fuzzy class is

$$C = S \cap T \cap L^2. \tag{17}$$

where C,S,T and L^2 are the respective non fuzzy classes associated to the properties s(x), t(x) y $l^2(x)$ respectively.

According with the statements the following predicates are the properties associated to fuzzy classes that represents Solid Economy Enterprises, Advanced Technology Enterprises, and very strong Line of Products Enterprises:

$$s(x)=f(x)\wedge v(x)\wedge(\neg(f(x))^{0.5}\rightarrow v^{2}(x)) \tag{18}$$

$$\begin{aligned} t(x) &= g(x)\wedge(p(x)\vee i(x)\vee d(x))\wedge \\ &\wedge(\neg g^{0,5}(x)\rightarrow(p^{2}(x)\vee i^{2}(x)\vee d^{2}(x)) \end{aligned} \tag{19}$$

$$l(x) = m(x)\wedge vl(x)\wedge ip(x) \tag{20}$$

The used predicates, like components in s, t and l are called basic predicates. All of them define fuzzy classes too.

Notice that 0.5 is used in the same way that 2 was used before, like a modifier expressing the word more or less.

Figure 2 illustrates the ontology through a logical tree. In the tree, conditional expressions are expressed, by placing on the arc the predicate corresponding to the premise; and as the end of the arc itself, the thesis.

The sigmoidal function of Fig. 3 illustrates the way to define the basic predicates and correspondent classes by membership functions.

Table 1 illustrates the answers that could be obtained requesting about if certain instances are part of each class defined in the ontology.

Compound predicates incorporated to the ontology using defined classes, defined connectives and modifiers are axioms. Axioms can be incorporated from human sources or by knowledge discovery.

For example the conditional s(x) → c(x), or the disjunction t(x) $\vee$ l(x) could be axioms. Equivalences of the form p(x)↔ a where p(x) is a predicate and "a" is a real number, could be very useful, using the possibilities of CFL to discover and to express knowledge from samples.

Table 1. Answer to fuzzy queries

Empresa x	f(x)	v(x)	g(x)	p(x)	i(x)	d(x)	ip(x)
A	0.5	0.47	0.3	0.93	0.81	0.61	0.6
B	0.6	0.63	0.5	0.41	1	0.95	0.8
C	0.9	0.75	0.7	0.62	0.55	0	1
D	0	0.99	0.8	0.81	0.79	0.7	0.5
	vl(x)	m(x)	s(x)	t(x)	l(x)	l2(x)	c(x)
A	0.23	0.1	0.5	0.516	0.234	0.058	0.246
B	0.77	0.4	0.611	0.682	0.627	0.393	0.545
C	0.92	0.8	0.812	0.584	0.903	0.815	0.728
D	0.39	1	0	0.763	0.58	0.336	0

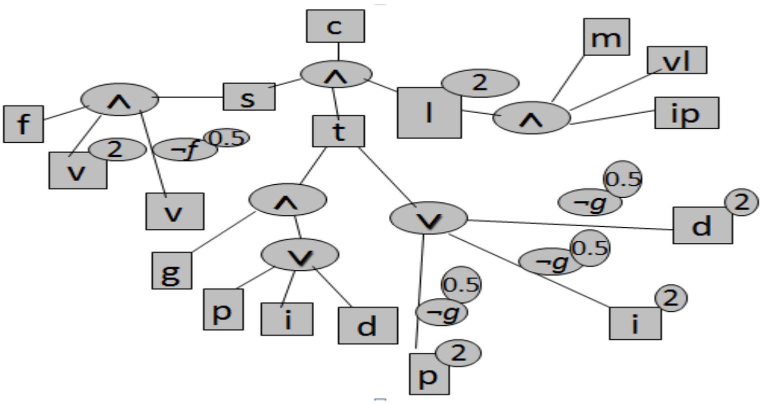


Fig. 2. Logical tree

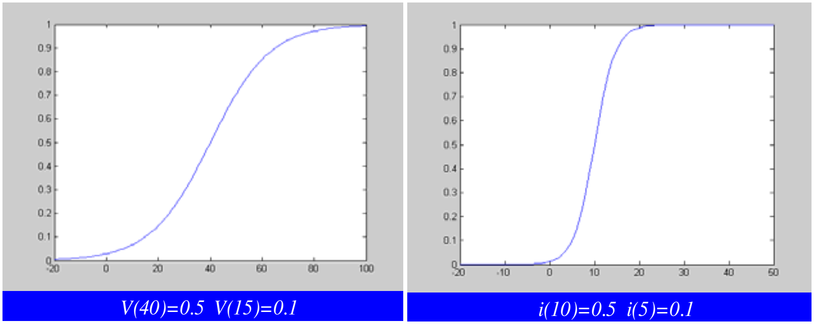


Fig. 3. Sigmoidal function

7 Conclusions

Properties of Compensatory Fuzzy Logic make possible a very useful representation of the knowledge: Compensatory Fuzzy Ontologies. They allow selection of relevant information, and useful knowledge discovery, very important for decision making.

Properties expressed by theorems 1 and 2 are especially important. Theorem 1 establishes that Compensatory Fuzzy Logic offers a way of reasoning completely compatible with the Boolean Logic.

Compensatory Fuzzy Ontologies can be used for querying and knowledge discovery from data and other properties, getting a way of automated ontology enrichment.

The fulfillment of Theorem 2 allows the use of the Compensatory Fuzzy Ontology to estimate real values of the universal proportions from samples, the compound inference associated with the theorem makes easy to work with a sample of instances to infer knowledge.

References

1. Friedman, J.: Estrategias Administrativas para la Eficiencia Universitaria. Revista Chilena de Administración Pública (6), 197–209 (2004)
2. Rajagopalan, N., Rasheed, A.M.A., Datta, D.K.: Strategic Decision Processes: Critical Review and Future Directions. Journal of Management 19(2), 349–384 (1993)
3. Espin, R., Fernández, E.: La Lógica Difusa Compensatoria: Una Plataforma para el Razonamiento y la Representación del Conocimiento en un Ambiente de Decisión Multicriterio. In: Análisis Multicriterio para la Toma de Decisiones: Métodos y Aplicaciones. Coedición: editorial Plaza y Valdes/editorial Universidad de Occidente (2009)
4. Tho, Q.T., Hui, S.C.: Automatic Fuzzy Ontology Generation for Semantic Web. IEEE Transactions on Knowledge and Data Engineering 18(6), 842–856 (2006)
5. Wallace, M., Avrithis, Y.: Fuzzy relational knowledge representation and context in the service of semantic information retrieval. In: IEEE International Conference, Budapest, pp. 1397–1402 (2004)
6. Fensel, D., van Harmelen, F., Horrocks, I., McGuinness, D.L., Patel-Schneider, P.F.: OIL: an ontology infrastructure for the semantic web. IEEE Intelligent Systems 16(2), 38–45 (2001)
7. Ghorbel, H., Bahri, A., Bouaziz, R.: A Framework for Fuzzy Ontology Models. In: Proc. of journées Francophones sur les Ontologies JFO 2008, France, pp. 21–30 (2008)
8. Zhai, J., Li, Y., Wang, Q., Lv, M.: Knowledge Sharing for Supply Chain Management Based on Fuzzy Ontology on the Semantic Web. In: International Symposiums on Information Processing (ISIP), pp. 429–433 (2008)
9. Dubois, D., Prade, H.: Fuzzy Sets and Systems: Theory and Applications. Academic Press Inc., London (1980)
10. Dubois, D., Prade, H.: A review of fuzzy set aggregation connectives. Information Sciences 36, 85–121 (1985)
11. Espin, R., Mazcorro, G., Fenández, E.: Consideraciones sobre el carácter normativo de la lógica difusa compensatoria. In: Infraestructura de Datos Espaciales en Iberoamérica y el Caribe. IDICT, Cuba (2007)
12. Espin, R., Fernández, E., Mazcorro, G., Marx-Gómez, J., Lecich, M.I.: Compensatory Logic: A fuzzy normative model for decision making. Investigación Operativa. Universidad de la Habana 27(2), 188–197 (2006)
13. Delgado, T., Delgado, M.: Evaluación del Índice de Alistamiento de IDES en Iberoamérica y el Caribe a partir de un modelo de Logica Difusa Compensatoria. In: Delgado, T., Crompvoets, J. (eds.) Infraestructura de Datos Espaciales: Iberoamérica y el Caribe. IDICT-CYTED (2007)

Conceptual Clustering and Analysis of Data from Gynecological Database

Cveta Martinovska

Faculty of Computer Science, University Goce Delcev, Krste Misirkov, bb
Stip, Macedonia
cveta.martinovska@ugd.edu.mk

Abstract. The aim of this work is to propose a methodology for classifying, analyzing and visualizing data of patients with different symptoms from gynecological database. The application implements a variant of WITT algorithm for conceptual clustering. Pre-clustering algorithm is proposed that includes a tradeoff between overlapping of the initial clusters and displacing the center of clusters far away from the region of great density. To overcome the problem with weak correlation different coding schemes for cases are tested. Successful approach was to take square root of attribute value intervals to achieve the intervals with different sizes. Two different datasets from gynecological database are used: data related to polycystic ovary syndrome and data relevant to diagnose pre-eclampsia.

Keywords: Conceptual clustering, incremental learning, data analysis, data visualization.

1 Introduction

Conceptual clustering is a type of unsupervised learning: object descriptions are used as input in the system and a classification scheme is produced [3]. As opposed to learning from examples conceptual clustering is learning by observation [1]. Conventional cluster algorithms lack descriptions of the clusters created. Conceptual clustering techniques, on the other hand, provide such descriptions.

The WITT algorithm for conceptual clustering is described by Hanson and Bauer [2]. The clusters created are accompanied by the correlational structure of the attribute values of the cases in those clusters. This correlational structure can be used to describe the clusters. Additionally, this structure enables determining the extent to which a new case belongs to each of the already defined clusters. This property also allows for incremental learning. Once a new case is added to a cluster, the correlational structure of that cluster may be updated. Talmon et al. [4] [5] found some undesirable properties of the measures that govern the clustering process and proposed some alternatives which overcome the problems they identified.

The organization of this paper is as follows. First, the WITT algorithm and corrections of clustering parameters are described. Next, the experimental results using gynecological database are presented. The paper ends with a short discussion of the results and conclusions.

2 WITT Algorithm and Corrections of the Clustering Parameters

WITT algorithm describes the created clusters using features of the members in a cluster. The main idea of this algorithm is to find clusters with members which have highly correlated feature sets that represent the clusters.

The measure that Hanson and Bauer propose for determining the cohesiveness between the existing clusters and a new cluster c is given by the formula

$$Cc = \frac{Wc}{Oc} \tag{1}$$

where Wc represents the within-cluster cohesion and Oc represents the average cohesion between c and all other clusters.

To compute Wc a contingency table is used. Contingency table Fij is a two - dimensional matrix where each element contains the number of cases with the same value l for attribute i and same value m for attribute j. The within-cluster cohesion Wc is defined as the average variance in the co-occurrences of all possible attribute-value pairs for a given cluster

$$Wc = \frac{\sum_{i=1}^{K-1}\sum_{j=i+1}^{K} D(i,j)}{\frac{K*(K-1)}{2}} . \tag{2}$$

In the above formula, D(i,j) is the co-occurence distribution which can be computed from the contingency table Fij, and K is the number of attributes. D(i,j) is defined as

$$D(i,j) = \frac{\sum_{l=1}^{L}\sum_{m=1}^{M} Fij(l,m)\ln Fij(l,m)}{\sum_{l=1}^{L}\sum_{m=1}^{M} Fij(l,m) * \ln(\sum_{l=1}^{L}\sum_{m=1}^{M} Fij(l,m))} \tag{3}$$

with L and M being the number of distinct values of attributes i and j respectively. L and M are equal to the number of rows and columns of the contingency table Fij.

Analyzing the D-measures Talmon et al. [4] [5], have noticed several misleading properties:

The value of D(i,j) is 1 when there is no variance at all in the co-occurences i.e. for each attribute all cases in the cluster have the same value. When the cases are evenly distributed in the contingency table, then the value of D(i,j) should be 0. In the contingency table with evenly distributed cases, when the number of cases is doubled, D(i,j) is greater and not the same, as was expected intuitively. So if the number of cases goes to ∞, D(i,j) goes to 1, which means that Wc also goes to 1. In the situation when the number of added cases in all clusters increases, it is difficult to distinguish these two situations. It is desirable that the size of contingency tables does not influence the value of D(i,j).

To overcome the observed problems, Talmon et al. [4] [5] propose using fraction of cases Fij= Fij /N, instead of Fij, where N is the number of cases in the cluster. So the entropy for the contingency table is computed as

$$E(i,j) = -\sum_{l=1}^{L}\sum_{m=1}^{M} FRij(l,m) * \ln FRij(l,m) \tag{4}$$

and it ranges from 0, when all cases are located in one element, to lnT when all cases are evenly distributed over the contingency table. T is the number of entries in a contingency table (T=L*M). The new formula for D-measure is obtained by normalizing E(i,j) with lnT and subtracting the result from 1

$$\delta(i,j) = 1 - \frac{\ln N - \dfrac{\sum_{l=1}^{L}\sum_{m=1}^{M} Fij(l,m) * \ln Fij(l,m)}{N}}{\ln T} \ . \tag{5}$$

Symbol δ is used for this new D-measure. The term Oc is the measure of the average cohesion of a cluster with all other clusters. It is defined as

$$Oc = \frac{\sum_{k=1,k\neq c}^{P} Bck}{P-1} \tag{6}$$

where P is the number of clusters or categories. Bck is relative cohesion between two clusters c and k. New measure βck proposed for Bck that is computed as

$$\beta ck = \frac{W^{*}c \cup k_{\max} - W^{*}c \cup k_{\min}}{W^{*}c \cup k_{\max} - Wc \cup k} \ . \tag{7}$$

With formula (7) βck is defined as the difference in within-cluster cohesion for the maximally similar and the disjunct union of the clusters, divided by the difference in the within-cluster cohesion for the maximally similar and the real union of the clusters. The union of two clusters results in a maximal value when as many as possible non-zero elements of the contingency tables of the first cluster are matched with non-zero elements of the contingency tables of the second cluster. On the contrary, the minimal value for the union of two clusters is obtained when as many as possible non-zero elements of the contingency tables of the first cluster are matched with zero elements of the contingency tables of the second cluster. Formula (7) shows that if clusters are maximally disjunct, the numerator and denominator are equal and βck has value 1, and if they are not disjunct βck is greater than 1. So Oc, which is an average of βck for a cluster c and all other existing clusters will vary between 1 and ∞.

In the original WITT algorithm the measure Cc is used to assess whether some case might be added in the existing cluster. Talmon et al. propose a normalized difference in within-cluster cohesion, rather then using the within-cluster cohesion of a cluster with one case added W_{c+1}

$$\gamma_c = \frac{\frac{W_{c+1} - W_{c+1\min}}{W_c - W_{c+1\min}}}{O_c} \quad . \tag{8}$$

Minimal within-cluster cohesion is achieved when the case that is added to a cluster is completely different from the cluster, so $W_{c+1\min}$ might be computed by

$$W_{c+1\min} = \frac{N_c W_c + 1}{N_c + 1} - \frac{(N_c + 1)\ln(N_c + 1) - N_c \ln N_c}{(N_c + 1)} * \frac{\sum \frac{1}{\ln T}}{N_T} \quad . \tag{9}$$

where Nc is the number of cases in a cluster c, and N_T is the number of contingency tables.

2.1 Pre-clustering and Refinement Algorithm

The conceptual clustering algorithm consists of two components: a pre-clustering algorithm and a refinement algorithm.

The basic idea of the pre-clustering algorithm is to have few initial clusters formed by pairs of cases, close enough to each other, and far away from other clusters. As a measure it takes the distance in a multidimensional space between the attribute values of the cases. First cluster is created from two cases that are most similar i.e. the distance between these two cases is smallest. After that, second cluster and other clusters take into account the distance between the centers of the new cluster and all of the existing clusters. The next cluster is created from two cases for which quotient d/Dist is minimal. The number of initial clusters N, depends on the previous information, regarding the number of expected groups. The main problem is how to code this information in the algorithm for the purpose of creating initial clusters.

The distance measure d for two cases i and j is defined as

$$d(i,j) = \sqrt{\sum_{k=1}^{K} (i[k] - j[k])^2} \tag{10}$$

where K is the number of attributes.

The other measure used in the previous description of the pre-clustering algorithm, Dist is computed as a sum of distances between the centers of each of the existing clusters and the new cluster

$$Dist = \sum_{c=1}^{C-1} DC[c] \tag{11}$$

where C is the number of existing clusters, including the new one. DC is the distance between the centers of an existing cluster and the new cluster, as can be seen from the following formula

$$DC[c] = \sum_{k=1}^{K}(center\_c[k] - center\_n[k])^2 \qquad (12)$$

where K is the number of attributes, and center_c [k] is an array with the coordinates of the cluster's center. In the above formula center_n is the center of the cluster that should be created. Coordinates of the cluster's center in the multidimensional space, taking into account that every initial cluster consists of two cases, can be computed as

$$center\_c[k] = \frac{i[k] + j[k]}{2} \qquad (13)$$

where i and j are two cases that form the cluster c.

The refinement algorithm adds cases to the initial clusters when those cases are similar enough to one of these clusters. When the refinement algorithm fails the pre-clustering algorithm tries to make new clusters using cases not yet assigned to existing clusters. When this process fails, existing clusters could be merged.

The algorithm does not always cluster all cases. When the remaining cases cannot be clustered and there are no clusters that can be merged, the algorithm stops.

3 Experimental Results

The application is used for clustering of two different datasets: data relevant to diagnose pre-eclampsia [6] and data related to polycystic ovary syndrome [7].

3.1 Clustering Data from Pre-eclampsia Database

Pre-eclampsia is a mediacal condition with pregnancy induced hypertension in association with significant amounts of protein in the urine. The database (see Table 1) consists of 154 cases with 3 attributes associated to protein excretion: amount of urinal protein, protein/creatinine ratio and activity of the urinary enzyme beta NAG. These parameters are measured for normotensive (with normal blood pressure) women in the first (22 cases), second (20 cases) and third (22 cases) semester of pregnancy and healthy women that are not pregnant (70 cases). The last group consists of 20 cases that represent women with pre-eclampsia.

Table 1. Sample of the pre-eclampsia database

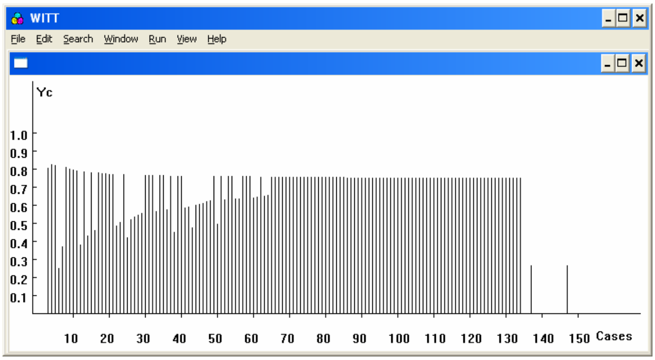
urinal protein (mg/l)	protein/creatinine ratio (mg/g)	NAG (U/g)
20	19.23	6.49
30	22.22	15.46
….	…	…
3020.2	1037.35	40.02
9600	8250	34.90

Several experiments are performed varying the number of initial clusters from five to two. Although five initial clusters are expected because of five different groups of cases experiments show overlapping of the clusters. Cluster of healthy not pregnant women overlaps with the three clusters of normotensive women.

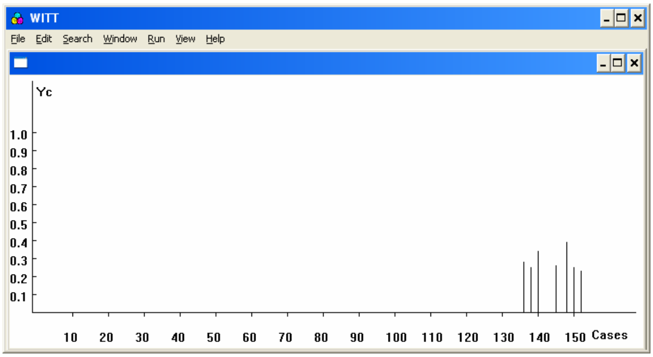
Results show that the clustering process is independent of the size of the clusters, because the cohesion measure takes into account the size of the contingency tables from which it is derived. The γ values for the initial clusters remain same in the situations when the initial clusters have one, two or ten cases. The last parameter that is examined is the number of distinct values for each attribute. Varying the number of values for attributes might change the number of cases that have the same value for different attributes and might help when cases are not well correlated.

The best results are obtained with 10 distinct values for each attribute and with two initial clusters. Linear coding scheme is used for the cases i.e. the intervals that represent distinct values for each attribute are same. Plots that represent results of the clustering process are shown in Fig. 1.

Cases with numbers from 1 to 64 belong to the groups of women in the first, second and third semester of pregnancy and from 65 to 134 to the control group. Cases that represent pregnant women with pre-eclampsia with numbers from 135 to 154 have large values for their attributes. The number of cases with the same large value is small and this leads to spreading these cases in different intervals.



a) Cluster of normotensive pregnant women and control group. Wc= 0.8874



b) Cluster of women with pre-eclampsia. Within-cluster cohesion is Wc= 0.5934.

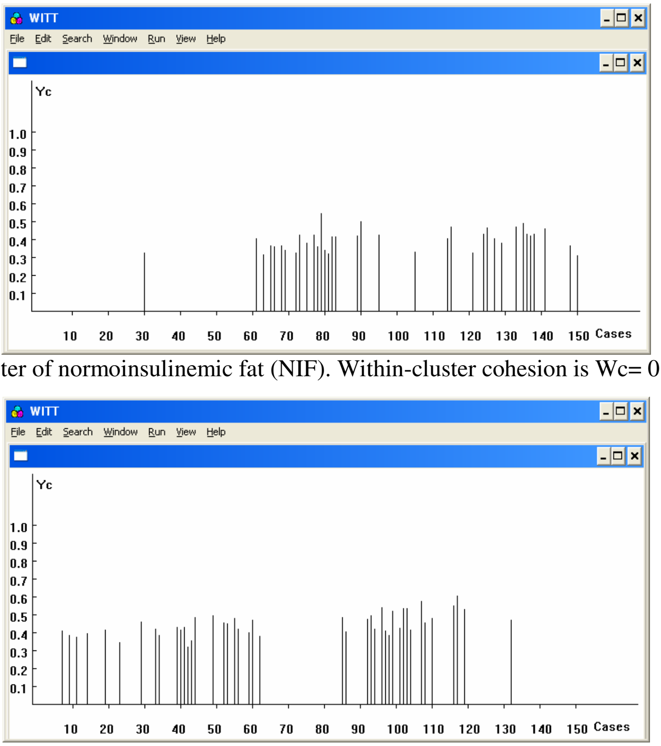
Fig. 1. Histograms with results of the clustering process. Threshold 1 is 0.4, the number of distinct values for attributes is 10 and coding of cases is linear.

3.2 Clustering Data from Polycystic Ovary Syndrome Database

Several experiments are performed using a polycystic ovary syndrome (POS) database. POS database is formed from the following groups of patients: normal, normoinsulinemic slim (NIS), normoinsulinemic fat (NIF), hyperinsulinemic slim (HIS), hyperinsulinemic fat (HIF). Each group consists of 30 patients.

In the database 14 attributes are stored for each patient but just the first 6 are used because their correlation structure is representative for the categories. These 14 attributes are: BMI - body mass index calculated as weight divided by square of height, WHR – waist/hip ratio, leptin, glycemia – concentration of glucose in the blood, insulin, G/I – glycemia/insulinemia ratio, TL – total lipids, TCh – total cholesterol, TG – triglycerides, HDL Chol – high density lipoprotein, LDL Chol – low density lipoprotein, F4 hormone, cortisol and age.

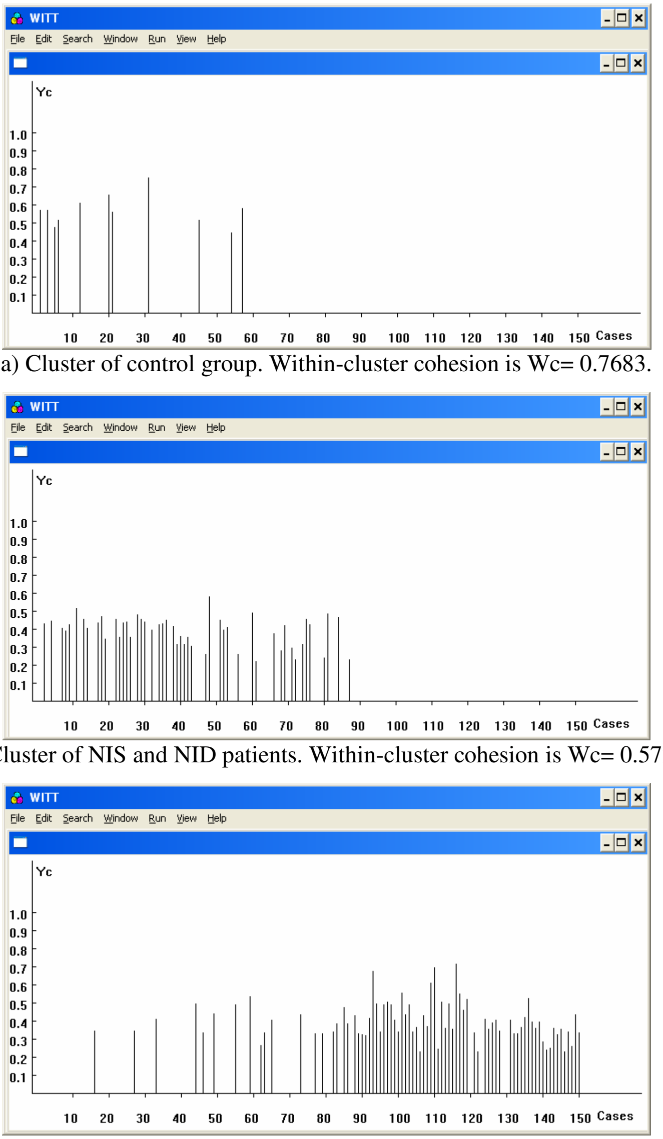
Age is dropped from the list of attributes because it does not contribute to the process of clustering. Attributes from 7 to 13 that form the lipid profile of the patients are also dropped. These attributes just put noise in the structure and are not relevant for the clustering process. They have been dropped after examination whether their absence will produce increase in the quotient Wc/Oc for all the clusters.



a) Cluster of normoinsulinemic fat (NIF). Within-cluster cohesion is Wc= 0.5581.

b) Cluster of hyperinsulinemic slim (HIS). Within-cluster cohesion is Wc= 0.7171.

Fig. 2. Histograms with results of the clustering process for POS databse. NIF cluster in addition to NIF cases contains HIF cases. HIS cluster besides HIS cases contains NIS cases.



a) Cluster of control group. Within-cluster cohesion is Wc= 0.7683.

b) Cluster of NIS and NID patients. Within-cluster cohesion is Wc= 0.5723.

c) Cluster of HIS and HID patients. Within-cluster cohesion is Wc= 0.4867.

Fig. 3. Histograms with results of the clustering process for POS databse with three initial clusters. Cluster of normals overlaps with the cluster formed from NIS and NID patients.

Plots in Fig. 2 are obtained with five initial clusters, number of distinct values for each attribute 10 and linear coding of value intervals. Only the cluster of normoinsulinemic fat (NIF) and hyperinsulinemic slim (HIS) patients are represented. Almost all of the patients with normal characteristics are in the cluster of normals. Some of the patients from the category normoinsulinemic slim (NIS) are in the HIS cluster, and some of the patients from the category hyperinsulinemic fat (HIF) are in the NIF. The antropometric parameters like BMI and WHR influence the clustering of these incorrectly classified cases that belong to other categories.

One reason that slim and fat normoinsulinemic patients are not clustered together might be that the cases in the initial clusters are not good representatives of these clusters. The same applies for clustering of the slim and fat hyperinsulinemic patients. Some alternative ways of creating initial clusters are considered. For example, the distance from formula (11) can be computed as

$$Dist = \sqrt{\sum_{c=1}^{C-1} DC[c]}\ . \tag{14}$$

With this change the distance between clusters is smaller, which can result in finding clusters that are not very far away from each other but they might be in the multidimensional space with greater density of cases.

Next experiment is performed using formula (14) for the distance Dist and starting the clustering process with three initial clusters. In this experiment a lot of cases remain outside of clusters showing that the cases were not well correlated. Different coding schemes were considered to overcome this problem. For example, the square root is taken from the intervals which represent the distinct values for the attributes, so that the first intervals are smaller and the last bigger. Other alternative that is examined is taking the same number of cases in each interval, and the third alternative was leaving the intervals with great number of cases (number of cases greater than square root of the number of cases in the data base divided by 2) the same and widening the intervals with smaller number of cases to have the 'necessary' number of cases. The best results are obtained when the square root is taken from the intervals.

Plots in Fig. 3 show the values of γ_c for cases in this experiment. Patients with normal values for the attributes are in the cluster of normals. The second cluster is constructed from the NIS and NID patients and the third from the HIS and HID patients.

4 Conclusion

The clustering process with the proposed alternatives for clustering measures is independent of the size of the clusters, because the cohesion measure incorporates the size of the contingency tables from which it is derived. The γ values for the initial clusters remain same when the initial clusters have one, two or ten cases.

Working with data from gynecological databases, several alternative ways for creating initial clusters are considered, because it is obvious that the clustering process is very much dependent on them. If there is overlapping among the initial clusters, the γ_c measures are very small, because the Oc measures are large. On the other hand, choosing the initial clusters that are far away from each other may result in finding the initial clusters which are not well correlated with the other cases that have to belong in the same cluster and they are not good representatives of their category. So the pre-clustering algorithm has to include a tradeoff between these two situations.

Visually, the ideal initial clusters have to be part of the multidimensional space where points are 'densely' packed, but to be far away from each other. If this requirement is not satisfied, for example if the region with great density for one of the categories does not exist, then the cases from that category are not clustered.

Another point to bear in mind is that there is prior information regarding the likely number of groups and this is helpful in finding the initial clusters. The main problem is how to include this information in the pre-clustering algorithm.

One useful suggestion for creating initial clusters is that the Oc measure has to be used to determine their number. As long as the Oc measures are 1 for every existing cluster c, they are acceptable as initial clusters, because they are disjunct.

References

1. Fisher, D.H.: Knowledge Acquisition via Incremental Conceptual Clustering. Machine Learning 2, 139–172 (1987)
2. Hanson, S.J., Bauer, M.: Conceptual Clustering, Categorization and Polymorphy. Machine Learning 3, 343–372 (1989)
3. Michalski, R.S.: Knowledge Acquisition through Conceptual Clustering: A theoretical Framework and Algorithm for Partitioning Data into Conjunctive Concepts. International Journal of Policy Analysis and Information Systems 4, 219–243 (1980)
4. Martinovska, C., Talmon, J.L.: Implementation of the alternative of the WITT algorithm. Techical report, University of Limburg, Faculty of Medical Informatics (1993)
5. Talmon, J.L., Braspenning, P.J., Fonteijn, H.: An analysis of the WITT algorithm. Machine Learning 11, 91–104 (1993)
6. Cekovska, S.: Biochemical Evaluation of the Functional Proteinuria. Ph.D. thesis. Faculty of Pharmacy – Skopje, Macedonia, pp. 87–94 (2006) (in Macedonian language)
7. Georgievska, J.: Influence of Insulin and Body Mass Index on the Endocrine and Metabolic Changes in Patients with Polycystic Ovary Syndrome. M.Sc. thesis. Faculty of Medicine. Clinic of Gynecology and Obstetrics (2006) (in Macedonian language)

System for Prediction of the Winner in a Sports Game

Eftim Zdravevski[1] and Andrea Kulakov[2]

[1] NI TEKNA – Intelligent Technologies, Negotino, Macedonia
eftim.zdravevski@ni-tekna.com
[2] University Ss. Cyril and Methodius,
Faculty of Electrical Engineering and Information Technologies, Skopje, Macedonia
kulak@feit.ukim.edu.mk

Abstract. This work presents a system that facilitates prediction of the winner in a sport game. The system consists of methods for: collection of data from the Internet for games in various sports, preprocessing of the acquired data, feature selection and model building. Many of the algorithms for prediction and classification implemented in Weka (Waikato Environment for Knowledge Analysis) have been tested for applicability for this kind of problems and a comparison of the results has been made.

Keywords: Data acquisition, data processing, decision-making, prediction methods.

1 Introduction

It is common knowledge that for many sports enormous amount of data is collected – for each player, team, game and season. Obviously this is too much data to be analyzed manually. This gave us the idea to test some algorithms for data mining on data sets that contain records of sport games. The data mining can be done from various aspects – prediction of final outcomes, prediction of player's injuries [8], prediction of future physical performances [7], discovering specific patterns (e.g. player B has made 60% of his field goals when player A was at point-guard position and has made 40% of his field goals when other point-guard was on the field [6]), as well as some other aspects. The goal of our research is to test various data mining algorithms for prediction of the final outcome (the winner) of a game. We don't aim to find out the exact reasons why a particular outcome was obtained, but to use a large set of outcomes to predict an unknown one. The classifiers that are used in the prediction process are implemented in Weka (Waikato Environment for Knowledge Analysis) [9].

A lot of research has been made in this area by experts who have the necessary domain knowledge for a particular sport, but also a solid background in mathematics. In many cases, they came up with complex formulas for particular type of performance in a game (offensive, defensive, etc.) and formulas for overall rating of players and teams [1] [2] [3] [4] [5]. The team rating formula can be very complex (it can contain more than 15 parameters), but also very important for the classification process. Sometimes the team ratings are used by some bookmakers to adjust the odds for a game.

Section 2 outlines the architecture of the proposed system. In section 3 each module of the system is described in more detail. The results of this research are presented in section 4 and a comparison of the results, obtained by different classifiers, is made.

2 System Design

In every data mining and knowledge discovery process the initial data has to go through few stages of processing in order to extract useful information. For this particular case of data mining in sports data, the stages of data processing are shown on Fig. 1.

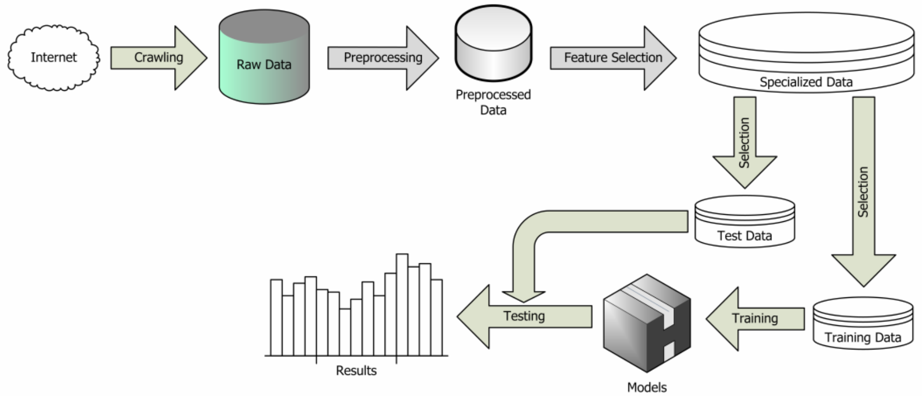


Fig. 1. The steps of data mining in records of sport games

The data processing in stages would be easier if the system is designed in a modular way. By doing that, each of the modules can be implemented and tested independently of the others, but also makes it easier to do modifications in one module without having to redesign the others. Each of the modules is dedicated to the processing of the information at a certain stage. There is a central module that integrates all specialized modules into a single system. Another good thing about this design is that other modules can be easily added to the system. That can be accomplished in the following way: first a new module would be designed and implemented and then the central module should be modified so it can use the new one. Such modules can contain implementations of algorithms for prediction or clustering that aren’t contained in WEKA. The modular design of the introduced system is shown on Fig. 2.

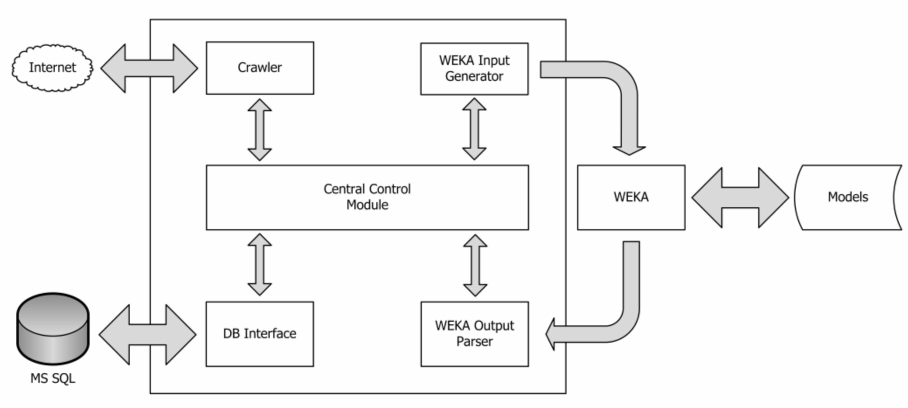


Fig. 2. The modular design of the system for outcome prediction

2.1 Data Collection

Obviously, to start testing algorithms for this problem a data set of games' records is needed. Although it is fairly easy to find the result and the statistics of a certain game on the Internet, to our knowledge there isn't any publicly available data set that can be downloaded and imported into some database. This enforces our system to have a module for acquiring information ("crawler") for the games of interest from the internet and storing it into a database.

2.2 Data Preprocessing

After all required data is stored in a relational database, it must be preprocessed. The preprocessing may refer to: normalization and/or discretization of some parameters in a given range; or generating new parameters that didn't exist in the original database. New parameters are generated by reviewing the data for the previous games of the current season. Previous games refer to games that were played before the date of the game whose data is being preprocessed and can contain data of games played by teams playing that particular game, but also games played by other teams. This means that none of the generated parameters uses "future" data, i.e. data that wasn't known before the beginning of a particular game. In other words, each generated parameter is time dependant and team dependant. Some of the parameters that are generated in this module of the system are:

- *Number of injured players in Team A before a particular game.* This parameter can be generated by reviewing the data from the previous game that Team A played, because it contains information why a particular player didn't play – either it was coach's decision or the player was injured. Additionally, there are websites that publish information about injured players on a daily bases. The information retrieval from this kind of websites can also be automated.
- *Winning streak (w) of Team A before a particular game.* This is done by counting how many games in a row have won (*w* is positive) or lost (*w* is negative) before that game.

- *Fatigue of Team A before a particular game.* We are introducing this parameter to indicate how many times did Team A have to travel in order to play the previous 7 games. Because the schedule of the games in NBA, WNBA, NHL and some other American sports is very busy (2-4 games per week), sometimes teams have to travel a lot in order to keep up with the schedule. Traveling a lot contributes to fatigue of the team. On the example shown on Fig. 3 the fatigue of Team A before game 27 (the particular game) is estimated to be 5/6 because it had to travel 5 times. The maximum fatigue is 1 (if the team traveled 6 times) and the minimum is 0 (if the team played all the relevant games at home).
- *Home, away and overall winning percentage.* The number of games won at home, divided by the number of games played at home, gives the home winning percentage. The calculation of the other parameters is similar.
- *Offensive, defensive and overall ratings of the team.* These ratings are calculated by formulas which are described in more details for various sports in [4] and for NBA basketball in [1] [2] [4].

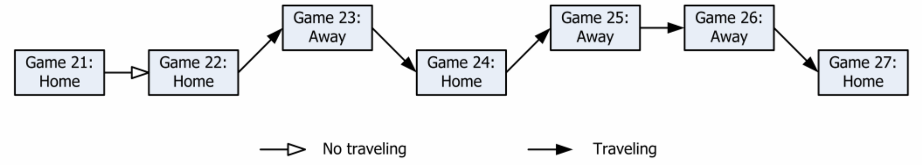


Fig. 3. Example of fatigue estimation

2.3 Feature Selection

Regardless of the domain of the problem and regardless of the classification or clustering algorithm, the training and test data have to be represented as a set of data points. Each data point is *N* dimensional space and each coordinate of the data point represents a feature.

The preprocessing phase enriches the set of available parameters for each game. However, it isn't practical to use all available parameters because it can lead to performance and precision degradation. It is important to determine which of the available parameters will be selected as features for the training and the test datasets. Some set of features may give better precision than other sets. The presented results in this research are obtained using 10 features. Some of these features are the ones described in the previous section. Another tweak that is done is grouping of two compatible parameters into one feature (e.g. instead of using as two different features the offensive rating of the home team and the offensive rating of the visiting team, their difference is used as a single feature). The names or the IDs of the teams that play a game aren't used as features of the game.

2.4 Training and Test Data Sets

For the purpose of this research we collected data for 2 consecutive NBA seasons from the official NBA website. This data contains detailed statistics of each game played during a season. The data from the first season is used as a training set and the data from the second season as a test set. There are 30 teams in the NBA league and each of them plays 82 games during a regular season, so a total of 1230 games are played. However, the first 20 games in a season of each team aren't considered neither for training nor testing, because they couldn't be represented by the features that we selected. Namely, in order to present a game as a data point to any prediction algorithm, it should be represented as a set of features. Some of the features that we decided to be most suitable for this problem need data from previous games (in the same season) and if we use the games from the beginning of a season then this data would be missing or would be incomplete. The following example shows why these games are avoided for training and testing: suppose we have a trained model and we want to predict the outcome of Game 6 of Team A. However, for a feature that corresponds to the average point margin of last 10 games we would need data from the previous 10 games (in the same season) of Team A and such data doesn't exist.

3 Implementation

In section 2 was given an overview of the design of the system and the purpose of each module and in this section their implementation will be discussed, conducted with the programming language C# using the .Net platform and a SQL Server database.

3.1 The Crawler

The task that the crawler performs is collecting data for games in a specific league and in a specific time period and inserting it in the SQL database. The data can be collected from the official website of the sport of interest where detailed statistics of many parameters are published.

Fortunately, the process of data collection from NHL, NFL, NBA, WNBA etc. can be automated. By manually examining the URLs where the final scores of games played on a particular date we have concluded that they have consistent format. Knowing that format, URL for any desired date can be automatically constructed. If there weren't any games on the specified date, then a web page for that date wouldn't exist or if it existed it would show a warning message. Either way, we would know that it doesn't contain information that is of any interest to us. Furthermore, the format in which the data is published on the webpage is also consistent – there are tables that contain the summary of the game and each player's accomplishments and they have a constant number of columns in a specific format. This enables HTML of the webpage to be parsed and the needed data to be stored in a database.

Everything mentioned here suggests that it is possible to develop an application that can fill in a database automatically for a given range of dates of games in a particular league. The collected data contains statistics of each player's and team's performance on each game. The crawler has to be specific to a particular sport and a particular league, since it uses its website to collect the needed data. Another limitation is that a major reconstruction of the webpage would imply that the crawler has to be modified as well. However, since our goal is to build a model from a data set of previous games to predict the outcome of future ones, we only need the data from few seasons. The data from the first one or two seasons can be used for training and the data from the following year for test and validation. The algorithms would be rated according to their precision on the test data set.

3.2 Preprocessing

Some of the preprocessing is done online while the data is being collected, because this way is more efficient. Most of the preprocessing methods are implemented as stored procedures and functions in the database. Some of them represent potential features, while others are just facilitating the computation of the former. The feature computation methods are invoked before the beginning of each training phase or test phase, meaning that they aren't invoked just once and their result stored in the database. Each time they are invoked their result is used as an input to the ARFF [9] generating module that prepares the input to the WEKA system. The results from the feature computation methods are not stored in the database for flexibility and scalability reasons. Namely, if the results are stored in the database, adding a new feature would entail redesign and update of the tables that store the results. There isn't such issue in the design we use. If a new feature is to be added, the function that computes it has to be implemented and invoked in the feature selection phase, which is far less complicated than the other possible solution.

3.3 Feature Selection

The feature selection is manual, i.e. we have to decide which features are to be taken into account. It is implemented as a stored function in the database that returns a table as a result. Each column in the resulting table represents a value of one feature, and each row represents a data point. This stored procedure takes as an input only two valid dates[1] and for each game played between those dates, a data point with the selected features is generated.

3.4 Interface to WEKA

In order to invoke classification, clustering or filtering algorithms from WEKA, an interface has to be implemented. WEKA algorithms can be invoked from the command line with a single command that has some specific parameters [9] – input ARFF file [9], model input/output file, algorithm name, etc. The ARFF files contain the input data set for the algorithm that is being invoked. They are generated using the results from the feature selection module. The output format from WEKA can be configured with the same command. The output has to be captured and then parsed so the parameters of our interest (e.g. predicted value) can be stored.

4 Results

In this section the results from our research are presented. The training and test data set contain data points corresponding to 930 NBA games each. The data that is in the

[1] Valid dates are dates from the regular season and dates that aren't in the beginning of the season for reasons explained in section 2.4.

training data set doesn't exist in the test data set. A referent classifier to which the others (implemented in WEKA) will be compared is a classifier that uses the following logic:

Let Team A (the home team) has rating *A*, and Team B (the visiting team) has a rating *B* before the beginning of a game that they are going to play. The rating is calculated using the Hollinger team rating formula [2]. If $A\text{-}B+3>0$, decide that this game will be won by Team A. Adding 3 in favor of Team A represents the home court advantage.

Table 1 shows the precision of the tested classifiers.

Table 1. Precision of the classifiers

Classifier	**Total Games**	**Correct**	**Incorrect**	**Precission**
functions_Logistic	930	677	253	0,728
meta_MultiClassClassifier	930	677	253	0,728
meta_ThresholdSelector	930	664	266	0,714
trees_NBTree	930	662	268	0,712
meta_RandomSubSpace	930	660	270	0,710
rules_JRip	930	658	272	0,708
functions_RBFNetwork	930	657	273	0,706
functions_VotedPerceptron	930	657	273	0,706
functions_SMO	930	651	279	0,700
trees_LMT	930	651	279	0,700
trees_ADTree	930	646	284	0,695
bayes_NaiveBayesUpdateable	930	646	284	0,695
meta_LogitBoost	930	646	284	0,695
meta_FilteredClassifier	930	644	286	0,692
bayes_NaiveBayes	930	644	286	0,692
meta_MultiBoostAB	930	641	289	0,689
meta_RandomCommittee	930	639	291	0,687
trees_RandomForest	930	639	291	0,687
trees_SimpleCart	930	639	291	0,687
trees_BFTree	930	632	298	0,680
bayes_BayesNet	930	632	298	0,680
Referent Classifier	**930**	**631**	**299**	**0,678**
meta_AdaBoostM1	930	629	301	0,676
rules_OneR	930	623	307	0,670

Table 1. (*continued*)

trees_REPTree	930	620	310	0,667
trees_DecisionStump	930	617	313	0,663
meta_Bagging	930	615	315	0,661
functions_MultilayerPerceptron	930	611	319	0,657
trees_J48	930	610	320	0,656
rules_NNge	930	608	322	0,654
misc_HyperPipes	930	596	334	0,641
meta_Stacking	930	592	338	0,637
rules_ZeroR	930	592	338	0,637
rules_PART	930	590	340	0,634
rules_ConjunctiveRule	930	583	347	0,627
trees_RandomTree	930	569	361	0,612
bayes_NaiveBayesSimple	930	482	448	0,518

The results show that the best classifiers have 5% better precision than the referent classifier which favors the team with better rating. They are 9 % better then the zero-R classifier that predicts the most common class (in this case the predicted winner is always the home team because it's the most common winner). Note that almost all of the classifiers from WEKA were used with their default settings. All of the classifiers in Table 1 are described in more detail in [9] and some of them in [10].

5 Conclusions and Future Work

This research showed that a system for prediction of the winner of a sports game can be designed and implemented. The precision it can provide is dependent on many parameters: the particular sport, the available data, the selected features, the classification algorithm, etc. Unfortunately, we have no base for comparison of our results. The referent classifier we define in section 4 uses greedy logic and we can't rely only on it. We couldn't find any set of predictions made by human expert or by some state-of-the-art artificial system for a complete season of some sport. If we could test our system on such set of games and compare our predictions to their predictions on the same set, then a better evaluation of our system could be made.

However, there are few things that can be done in order to improve the precision of the predictions. One thing that we can do is first to cluster the training and test data sets and then use a different model for each cluster. The logic behind this idea is that some teams rarely lose many games in a streak, while others rarely win many games in a streak. There is no guarantee that this modification will contribute to more precise predictions, but it's something worth trying. Another thing that can be tried is to use aggregation of different classifiers in order to improve the degree of belief of some predictions or to improve the overall precision of all predictions. As it was mentioned earlier, the feature selection is manual. This phase can be modified by automating it, so different combination of features can be tested. Such modification may contribute to better results because the human factor in the feature selection would be removed.

References

1. Oliver, D.: Basketball on Paper: Rules and Tools for Performance Analysis. Potomac Books (2005)
2. Hollinger, J.: Pro Basketball Prospectus. Potomac Books (2002)
3. Basketball terms and formulas, `http://www.basketballreference.com`
4. APBRmetrics, `http://en.wikipedia.org/wiki/APBRmetrics`
5. Albert, J., Koning, R.H.: Statistical thinking in sports. Chapman & Hall/CRC, Boca Raton (2008)
6. Bhandari, I., et al.: Advanced Scout: Data Mining and Knowledge Discovery in NBA Data. In: Data Mining and Knowledge Discovery, vol. 1, pp. 121–125. Kluwer Academic Publishers, Dordrecht (1997)
7. Fieltz, L., Scott, D.: Prediction of Physical Performance Using Data Mining. Research Quarterly for Exercise and Sport 74 il, A-25 (2003)
8. Flinders, K.: Football Injuries are Rocket Science (2002.10.14), `http://www.vnunet.com/vnunet/news/2120386/football-injuries-rocketscience`
9. Witten, I.H., Frank, E.: Data Mining: Practical machine learning tools and techniques, 2nd edn. Morgan Kaufmann, San Francisco (2005)
10. Duda, R.O., Hart, P.E., Stork, D.G.: Pattern Classification, 2nd edn. John Wiley & Sons, Inc., Chichester (2001)

Computer Generated News Site – TIME.mk

Igor Trajkovski

Faculty of Computer Science and Information Technology,
New York University Skopje, 1000 Skopje, Macedonia
admin@time.mk
http://www.time.mk/trajkovski/

Abstract. The Internet has broken down the barriers that exist between people and information, effectively democratizing access to human knowledge. Nowhere is this more apparent than in the world of news. According to a recent survey news browsing and searching is one of the most important internet activity. The huge amount of news available on line reflects the users need for a plurality of information and opinions. We believe that more information means more choice, more freedom and ultimately more power for people. TIME.mk is our attempt to connect users with the most important current news stories. It pulls together the most reported stories on its front page so that users do not have to examine the web for up-to-date stories. In this paper we present the architecture of TIME.mk and describe all the details of its functioning.

Keywords: news engine, information extraction, text similarity, clustering, classification.

1 Introduction

According to a recent survey [1] made by Nielsen/NetRatings for Newspaper Association of America, news browsing and searching is one of the most important internet activity. In June 2009, there were more than 70.3 millions of active news users in U.S.. The huge amount of news available on line reflects the users need for a plurality of information and opinions. News Engines are then a direct link to fresh and unfiltered stream of information. There are many commercial news engines. Google News retrieves news information by more than 4,500 sources, organizes it in categories and automatically builds a Web page with the most important news for each category. Yahoo news runs an analogous service on more than 5,000 sources. A list of commercial news engine is given in [2].

Despite this great variety of commercial solutions, we found just few academic papers on this subject. NewsInEssence [3] is a system for finding and summarizing clusters of related news articles. Chung [4] proposes a topic mining framework for news data stream. Henzinger [5] finds news articles on the web that are relevant to TV news currently being broadcast. Reis [6] proposes a tool to automatically extracting news from Web sites. We think that the few scientific publications cause the news engine technology to remain largely a secret art.

2 TIME.mk as a News Engine

In this paper, we introduce a general framework to build a News Engine. Our system, TIME.mk, is a complete news engine for retrieving, indexing, clustering, classifying, ranking and delivering news information extracted both form the Web and from news feeds. Currently TIME.mk collect news from 50 Macedonian news sources (TV stations, newspapers, news agencies, etc. and every day we get more requests for inclusion), groups similar articles together and displays them in order of their calculated importance.

By linking people to so many news sources, TIME.mk makes it much easier for them to read about stories from different angles and to search for more information on issues of their interest. The strength of this project is the respect of copyright. TIME.mk never shows more than the headline, a snippet and a thumbnail image of the news article. If people want to read the entire news story or see the full photo, they have to click through to the news source's website.

TIME.mk is similar by function to the Google's search/news engine in that it collects all the news it can find, creates an index of that information, and serves it to the users via a simple web interface. Our goal is to give users the most relevant, objective results, which is why we generate them automatically and without human intervention.

TIME.mk organizes articles so that many texts from different sources of a single news story appear in a group. We call these 'news clusters'. This approach groups headlines from different publications together, providing users with multiple viewpoints on a given news story. Publishers often ask us how we decide which clusters and type of news appears on the TIME.mk homepage. The short answer is: we don't decide.

The headlines on the TIME.mk homepage are selected entirely by an algorithm, based on many factors including how often and on which sites a news story appears elsewhere on the web. Basically, we look at the number of original articles being produced and published by editors in order to determine the size of a news story, which we also weight based on how recent it is.

Say, for example, that TIME.mk registers that in a one-hour period of time a cluster of two news articles about a football match in Germany appears; whereas a news story about EU-Serbia-Kosovo talks nets 20 articles. The algorithm detects that the latter is a bigger news story and gives that cluster priority in the ranking.

The system is made up by the modules depicted in Fig. 1 and has a Web interface at `http://www.time.mk/`. In the following, we describe them.

2.1 Retrievers

The task of this module is gathering links of news articles and extracting text and images from the web pages of these news articles.

TIME.mk uses two methods for links gathering. One is with RSS and another with regular expression(s) directly from the web pages of the news sources. In

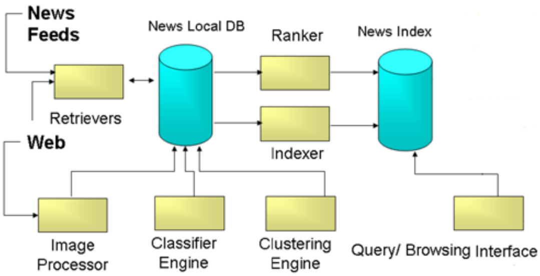


Fig. 1. Architecture of TIME.mk

the second method, for each news source we have to write regular expression(s) for detecting the relevant links.

Text extraction is the problem of selecting text and image on a web page that represents the content of the news. TIME.mk uses two methods for solving this problem. The first method sees news artislec's web page as simple array of characters, which is translated into array of numbers, one number for each character. If the character is a part of a HTML tag, then it is translated into a negative number, othervise it is translated into a positive number. Then the problem of text extraction is transformed into the search of continuous subarray with maximal subsum. The second method first creates DOM (Document Object Model) tree of the HTML data. Then it extracts all the text that is found inside news source specific set of tags (nodes).

Table 1. Comparision of the text extraction methods

Method	Advantages	Disadvantages
Maximal subsum	General, it works for all news sources	not suittable for short texts, it can extract text that is not part of the news
DOM tree	Accurate	requires news site specific html tags

If links of the news articles are not gathered by RSS, for each source we need to provide urls where links of the news articles can be found. These urls are called 'hubs'. By providing the hubs for given news source, we also provide the category of the news articles found at these hubs. Thus, the news articles enter the system as already classified (by category) and can be used as clean data for training the classifier that is used for classifying news articles that are not categorized.

2.2 Keywords Extraction

Keyword extraction process is used after the text of the news articles is extracted and its task is to extract single words (terms) characteristic for the news article. Extracted keywords are used for classification and clustering of the news articles.

For solving this task we used vector space model [7] for representing news articles. A document is represented as a vector and each coordinate of the vector corresponds to a separate term. If a term occurs in the document, its value in the vector is non-zero. Several different ways of computing these values, also known as term weights, have been developed [7]. The dimensionality of the vectors is the number of words in the vocabulary (the number of distinct words occurring in all news articles).

In the classic vector space model proposed in [7] the term specific weights in the document vectors are products of local and global parameters. The model is known as Term Frequency-Inverse Document Frequency (TF-IDF) model. The weight vector for document d is:

$$v_d = [w_{1,d}, w_{2,d}, ..., w_{K,d}]^T$$

where:

$$w_{t,d} = tf_t \cdot \log \frac{|D|}{|D_t|}$$

and:

- tf_t is term frequency of term t in document d (a local parameter)
- $\log \frac{|D|}{|D_t|}$ is inverse document frequency (a global parameter), where $|D|$ is the total number of news articles; $|D_t|$ is the number of news articles containing the term t and K is the number of distinct words occurring in all news articles.

After the term weights are calculated the news article is represented with the top N_{kw} terms (keywords) and its appropriate weights. At the end this vector is normalized ($\|v_d\| = 1$).

One applications of this kind of representation is the computation of news articles similarity, by calculating the angle between news articles vectors. In practice, it is easier to calculate the cosine of the angle between the vectors instead of the angle, that is why it is called *cosine similarity*:

$$\cos\theta = \frac{v_1 \cdot v_2}{\|v_1\|\|v_2\|}$$

A cosine value of zero means that the news articles are orthogonal and have no major similarities.

2.3 Image Processor

This module analyzes information stored in the news database and tries to enrich each news article with an associated image. In the easy case the news source has already associated an image to a given news in the RSS feed. In other cases, we have a news n, extracted by a Web page p or by a RSS feed, with no associated image which refers to a Web page p. In this case we use heuristic to identify the most suitable image to be associated to n, from other news articles that have an image. We take the image from the n's most similar news article (that has image) in the news database.

2.4 Classifier Engine

All not classified news articles needs to be classified. The categories in the system are predefined and given in Table 2. Note that a (relatively large) part of the RSS feeds and hub pages are already classified from the originating news source. As a consequence, the key idea for classifying is to use the classifier in a mixed mode: as soon as an already classified news by a news source is seen, the classifier is switched in training mode; the remaining unclassified news are categorized with the classifier in categorizing mode. We use a naive Bayes classifier [8]. The probability of a given document d composed of words $w_1, w_2, ..., w_n$, belonging to a class C, is:

$$p(C|w_1, w_2, ..., w_n) = \frac{p(C)\ p(w_1, w_2, ..., w_n|C)}{p(w_1, w_2, ..., w_n)}$$

In practice we are only interested in the numerator of that fraction, since the denominator does not depend on C and the values of the words w_i are given, so that the denominator is effectively constant. Using the "naive" conditional independence of occurance of the words w_i, the numerator is equivalent to the joint probability model:

$$p(C, w_1, w_2, ..., w_n) = p(C) \prod_{i}^{n} p(w_i|C)$$

where: $p(C)$ is the fraction of article belonging to C, and $p(w_i|C)$ is the probability that document beloning to C will contain the word w_i. Both parameters are easily computed.

Finally, the corresponding classifier is the function *classify* defined as follows:

$$\mathit{classify}(d) = \mathit{classify}(w_1, w_2, ..., w_n) = \mathit{argmax}_c\ p(C = c) \prod_{i}^{n} p(w_i|C)$$

2.5 Clustering Engine

Clustering engine has a task to group together news articles that report about the same event. These groups are called news stories. We don't know in advance the exact number of news stories and their coverage (number of articles covering a single story). Therefore we use hierarchical agglomerative clustering (HAC) for solving this task [8]. HAC treat each news article as a singleton cluster at the beggining and then successively merge (or *agglomerate*) pairs of most similar clusters until all clusters have been merged into a single cluster that contains all news articles. A HAC clustering is typically visualized as a dendrogram as shown in Fig. 2.

HAC does not require a prespecified number of clusters. However, we want a partition of disjoint clusters. In this case, the hierarchy needs to be cut at some point. We cut at a prespecified level of similarity. For example, we cut the dendrogram at 0.4 if we want clusters with a minimum similarity of 0.4.

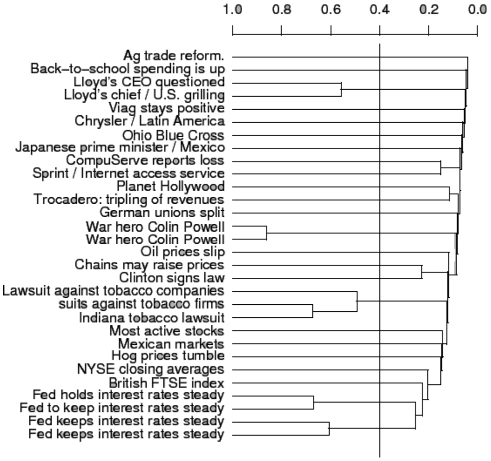


Fig. 2. A dendrogram of a HAC clustering of 30 documents is shown. Cut of the dendrogram at 0.4 is shown. 24 clusters are formed. Borrowed from [8].

2.5.1 Time Complexity of HAC

The naive implementation of HAC is $O(n^3)$. We need $O(n)$ steps of selecting the most similar clusters, mergin these two clusters and recalculating the similarity of the new cluster with the remaining ones. Naive selection of the most similar clusters is $O(n^2)$. We used centroid clustering for calculating the similarity between clusters which give us $O(1)$ time for merging two clusters. All these steps in total give $O(n^3)$ complexity, which is quite slow if we try to cluster thousands of news articles. But if we use priority queue (PQ) for storing the similarity matrix then the total complexity will be $O(n^2 \log(n))$. Keep in mind that inserting and deleting an element in PQ is $O(\log(n))$ and selecting the minimum is $O(1)$.

2.5.2 Classification of Clusters

After the news stories are created as a result of the news article clustering, we need to classify each news story into one of the ten categories. This classification is needed becouse we can not expect that all articles in one cluster will be from one category. For example, event about the world economic crisis one source can publish it in the world section, and another source can publish it in economy section. We solve this problem by simple majority voting. If two categories have the same number of votes we clasify the article with the more specific category (Balkan is more specific than World, Economy is more specific than Macedonia and Word, etc.).

2.6 Ranker

Ranking news articles is a rather different task than ranking Web pages. From one side, we can expect a smaller amount of spam since news stories come from controlled sources. When a news article is issued, we can have two different scenarios: the news article can be completely independent on the already published stories, or can be aggregated to a (set of) news articles previously posted. Anyway, we stress that, by definition, a news article is a fresh piece of information. For this reason, when a news article is posted there is almost no HTML link pointing to it. Therefore, HTML link based analysis techniques, such as PageRank, can produce a limited benefit for news ranking.

Any ranking algorithm for news stories (clusters) should have at least the following four properties:

- *Time awareness.* The importance of a piece of news changes over the time. We are dealing with a stream of information where a fresh news story should be considered more important than an old one.
- *Important News articles are Clustered.* An important news story is probably (partially) replicated by many sources. From the news engine point of view, this means that the (weighted) size of the cluster is a measure of its importance.
- *Authority of the sources.* The algorithm should be able to assign different importance to different news sources according to the importance of the news articles they produce. So that, a piece of news coming from "BBC" can be more authoritative than a similar article coming from say "Kirilica", since "BBC" is known for producing good stories.
- *Diversity.* News story reported by big number of sources should be more important than news story reported by small number of sources.

This is the formula used by TIME.mk for calculating the weight of the cluster c at moment t:

$$WC(c,t) = SourceEntropy(c) \cdot \sum_{i=1}^{k} WN(n_i,t)$$

where:

- k is the size of the cluster c. n_i, $(1 <= i <= k)$ are the news articles composing the cluster c.
- $SourceEntropy(c)$ represents the entropy of the set of news sources that are included in the cluster. This value is scaled and its value range is from 1 to 2. If all sources in the cluster are different $SourceEntropy(c) = 2$. If $k > 1$ and all sources in the cluster are the same then $SourceEntropy(c) = 1$.
- $WN(n_i,t)$ the weigth of news article n_i at time t which has been published at time t_i:

  $$WN(n_i,t) = A(source(n_i)) \cdot e^{-\alpha(t-t_i)}, \quad t > t_i$$

  The value α, which accounts for the decay of "freshness" of the news story, is obtained from the half-life decay time θ, that is the time required by the rank to halfe its value, with the relation $e^{-\alpha\theta} = \frac{1}{2}$. This α can depend on the category to which the news article belongs. For instance, it is usually a good idea to consider 'Sport' news decaying more rapidly than 'Fun/Showbiz' news.

- $A(s)$ accounts for the authority of the source. One source is more authoritive if it is more cited than other sources. Every time when our system detects a duplicate of a news article, published by two or more different sources, the source that first published the article gets a credit. Duplicate news articles are detected by n-gram method.

The final step is sorting the clusters according to their weight and creation of the static HTML files that are served to the users.

3 Implementation Details

The news engine is running on a single PC with a Pentium IV 3GHz, 2GB of main memory. It is completly implemented in programming language Python.

Currently, we selected a list of 50 news sources (most popular TV stations, newspapers, news agencies, etc.). For efficiency reason, the space of news sources is partitioned and this module is composed by several processes which run in parallel. The data is collected 24h per day, updated every 10 minutes - 144 times a day, and the stream of information is stored into a news database.

The complete cycle of crawling the all 50 sources, extrcting the text from the new news articles, clustering the news articles and scoring the news articles and news stories, takes approximatelly 3 minutes (depending on the response time of the web sites of the news sources). Clustering is performed only on news aritcles published in the last 2 days, which gives us approximatelly 2500 news articles. This number of news articles is clustered in less than one minute.

Number of keywords, N_{kw}, was determined experimentaly and it has a value of 7. We don't take all the words in a news article in its representation, becouse most of them have small weight compared to the 5-6 most important keywords. Also clustering algorithm is running faster becouse the time complexity of the cluster similarity function is linear in the dimensionality of the representation vectors. We can not chose less keywords, becouse small number of keywords can not accurately represent the content of the news article.

Also, not all news sources have the same categories as those predefined in our system. Therefore we need to map news source's categories into TIME.mk categories, if such a mapping is possible. Othervise news articles are left uncategorized and we leave classifier to categorize them. In some cases we help the classifier, by suggesting that news article is belonging into one of the two presumed classes, so classifier need to decide between two classes, not between ten classes. For example, if some source has not category Balkan, it publishes its article about Balkan in section World. We suggest to the classifier that news articles from this hub belong to the category Balkan or World. Now classifier has much easier job to do (the classification error is smaller).

4 Results

For space reason, we report just the most important results. For evaluating the quality of the results, we used the data set collected by TIME.mk, gathering news articles from more than 50 continuously updated sources. The data set consists of about 300,000 news articles collected over a period of one year (from 01/07/08 to 30/06/09) and classified in 10 different categories (see Table 2).

Table 3 and Table 4 present the precision and recall data of the two classification problems. In the first problem, if the news articles was labeled with two categories, the classification was considered correct if classifier classified the article in one of these two categories. For the second problem we used randomlly choosed 100 (per category) manually labeled clusters as an evaluation set. Selected clusters had minimum two news articles. As we can see the precision and recall numbers are hight, becouse the probability that more than half of the articles to be incorrectly classified is very low. Most of the rare errors, come from classification of small clusters (size two or three) that rearly come on the homepage, so basically on the homepage we have more than 99% accuracy.

Table 2. Number of news articles, per category, in a data set

Category	#news	**Category**	#news
Macedonia	58,596	Sport	43,498
Balkan	19,341	Chronicle	10,108
World	29,647	Culture	16,933
Economy	29,754	Technology	5,755
Skopje	7,325	Fun/Showbiz	37,798
Uncategorized	58205	**Total**	316960

Table 3. Precision/Recall of the Naive Bayes classifier for classifying articles.

Category	Precision	Recall	**Category**	Precision	Recall
Macedonia	84%	90%	Sport	92%	86%
Balkan	82%	81%	Chronicle	80%	77%
World	90%	87%	Culture	78%	82%
Economy	78%	80%	Technology	91%	94%
Skopje	73%	62%	Fun/Showbiz	89%	88%

Table 4. Precision/Recall of the majority voting algorithm for classifying clusters.

Category	Precision	Recall	**Category**	Precision	Recall
Macedonia	95%	98%	Sport	98%	96%
Balkan	94%	96%	Chronicle	98%	98%
World	98%	98%	Culture	98%	96%
Economy	97%	96%	Technology	98%	98%
Skopje	98%	96%	Fun/Showbiz	98%	97%

5 Conclusion and Future Work

In this paper we have presented an implementation details of a full scale news engine. Our work has been motivated by the large usage of news engines versus the lack of academic papers in this area. An extensive testing on more than 300,000 news articles, posted by 50 sources over one year, has been performed, showing very encouraging results.

The future work on the engine will be focused on scoring the clasters. At the moment the score of the news articles does not depent from the amount of new information that they introduce. For example if new news article is published and is clustered in a news story, most probably it will be ranked first in its cluster (except if it is duplicate), even if it does not introduce any new information. The idea is to include amount of new facts, that one article brings to the news story, into the score of the news article. This will require a more advanced semantic model for representing news, compared to the current "bag of words" model.

References

1. http://www.naa.org/PressCenter/SearchPressReleases/2009/NEWSPAPER-WEB-SITES-ATTRACT-MORE-THAN-70-MILLION-VISITORS.aspx
2. http://searchenginewatch.com/
3. Radev, D., et al.: NewsInEssence - a system for domain-independent, real-time news clustering and multi-document summarization. In: Proc. of the First Int. Conf. on HLT Research, San Diego, March 18-21, pp. 1–4 (2001)
4. Chung, S., McLeod, D.: Dynamic topic mining from news stream data (2003)
5. Henzinger, M., et al.: Query-Free News Search. In: Proceedings of the 12th International WWW Conference (2003)
6. Reis, D., et al.: Automatic Web News Extraction Using Tree Edit Distance. In: Proc. of 13th WWW Conference (2004)
7. Salton, G., et al.: A Vector Space Model for Automatic Indexing. Communications of the ACM 18(11), 613–620 (1975)
8. Manning, C., et al.: Introduction to Information Retrieval. Cambridge Press (2008)

Applying Bagging Techniques to the SA Tabu Miner Rule Induction Algorithm

Ivan Chorbev[1] and Mirjana Andovska[2]

[1] Faculty of Electrical Engineering and Information Technologies,
P.O. Box 574, MK-1001 Skopje, Republic of Macedonia
ivan@feit.ukim.edu.mk
[2] Netcetera. ul. Partizanski Odredi 72a, MK-1000 Skopje, Republic of Macedonia
m.andovska@gmail.com

Abstract. This paper presents an implementation of bagging techniques over the heuristic algorithm for induction of classification rules called SA Tabu Miner (Simulated Annealing and Tabu Search data miner). The goal was to achieve better predictive accuracy of the derived classification rules. Bagging (Bootstrap aggregating) is an ensemble method that has attracted a lot of attention, both experimentally, since it behaves well on noisy datasets, and theoretically, because of its simplicity. In this paper we present the experimental results of various bagging versions of the SA Tabu Miner algorithm. The SA Tabu Miner algorithm is inspired by both research on heuristic optimization algorithms and rule induction data mining concepts and principles. Several bootstrap methodologies were applied to SA Tabu Miner, including reducing repetition of instances, forcing repetition of instances not to exceed two, using different percentages of the original basic training set. Various experimental approaches and parameters yielded different results on the compared datasets.

Keywords: Bagging, Bootstrap, Simulated Annealing, SA Tabu Miner, Tabu Search, Data Mining, Rule Induction.

1 Introduction

The SA Tabu Miner (Simulated Annealing and Tabu Search based Data Miner) [1] is developed to perform the classification task of data mining in an integrated medical expert system. It is a rule induction algorithm which creates rules incrementally, performing a sequential process to discover a list of classification rules to cover as many training cases as possible with the highest quality. It uses a combination of Simulated Annealing and Tabu Search to perform the search for the optimal classification rule. It has been compared with CN2 and Ant miner algorithms on public domain data sets. The results showed that it obtained similar and often better results than the other approaches [1].

A major concern with a rule induction algorithm is how to improve its predictability. One possible solution is to use ensemble methods [2, 3, 4]. The ensemble methods were designed to improve the predictive accuracy of various learning algorithms. Given a set *S* of training examples, a learning algorithm outputs a classifier, which is a hypothesis about the true function *f*. An ensemble of classifiers is a set of individually trained classifiers whose predictions are combined in some way (typically by weighted or unweighted voting) to classify new examples. The resulting classifier is often much more accurate than any of the single classifiers consisted in the ensemble.

Many methods for constructing classifier ensembles have been developed [3]. One group of methods manipulates the training examples by using re-sampling techniques. These methods include bagging [3], boosting [7] and cross-validation committees [11]. Boosting generates the classifiers sequentially, increasing the weights of the misclassified examples which are provided as input to the next classifier, while the other two methods generate the classifiers in parallel. In this paper, we focus our attention towards the bagging method introduced by Brieman [5]. Inspired by its simplicity and motivated by its rare application in rule induction algorithms, we applied bagging with several bootstrap methodologies on the rule induction SA Tabu Miner algorithm, and obtained different results on different datasets.

The paper is organized as follows: in the next section we describe the bagging approach. The implementation of bagging over the rule induction SA Tabu Miner algorithm is given in section 3. Section 4 covers our experimental results. The conclusions are presented in the final section.

2 Bagging

Bagging uses bootstrap sampling, which is sampling with replacement, to manipulate the input data in order to get several different versions of learning sets. Then, the learning algorithm is applied to the learning sets in order to obtain a set of diverse classifiers. Finally, the resulting multiple classifiers are combined by majority voting to form a composite classifier. The classifier trained on trail *i* is denoted as C_i, while C^* is the bagged (the final, composite) classifier.

At the time of its invention, only heuristic arguments were presented to explain why bagging would work. Breiman [3] presented empirical evidence that bagging can indeed reduce prediction error. He applied bagging over seven medium sized datasets and reported results that when the number of bootstrap replicates *T* is set at 50, the average error of the bagged classifier C^* ranges between 0.57 and 0.94 of the corresponding error when a single classifier is learned. In context of the number of bootstrap replicates, he examined 10 to 100 replicates and suggested that fewer replicates should be involved when the output is numerical. He also introduced the concept of an order-correct classifier-learning system as one which, over many training sets, tends to predict the correct class of the test instance more frequently than any other class. He explained that aggregating classifiers produced by an order-correct learner results in an optimal classifier, even though a single order-correct learner may not produce the optimal classifier. He also demonstrated that bagging is most effective for cart trees, due to their instability. Accordingly, he showed that bagging works especially well for unstable learning algorithms – algorithms where with small changes in the training set the resulting classifier experiences large changes in its prediction. Decision-tree, neural network, and rule learning algorithms are all unstable. Linear regression, nearest neighbor, and linear threshold algorithms are generally very stable [3].

Recently bagging has attracted a lot of attention, both experimentally, since it behaves well on noisy datasets, and theoretically, due to its simplicity and also due to the popularity of the bootstrap methodology, which is an essential characteristic of bagging. Consequently, new researches have been conducted to explain the potential influence of training examples in the prediction process. In [9] experimental evidence is presented to support the hypothesis that bagging stabilizes prediction by equalizing the influence of training examples. Hence, influence equalization is mainly due to the absence of influential examples in 37% of bootstrap samples.

On the other hand, Buja and Stuetzie [18] observed bagging U-statistics and reported nearly identical results when using bootstrap sampling with or without replacement, in terms of bias, variance, and MSE, for resample size $M_{w/o}$ if $N/M_{with} = N/M_{w/o} - 1 = g$ ($>0, <\infty$), where N is the size of the original training set, and M_{with} and $M_{w/o}$ the resample size of the training set when using re-sampling with and without, respectively. Hall and Samworth [19] study the properties of bagged nearest neighbor classifiers and address how performance depends on the resample size. It is interesting to note that the benefits of the random sampling could be applied to other ensemble algorithms. For example in [10], Friedman proposes to incorporate uniform random sub-sampling, at each stage of the gradient boosting algorithm. Buhlmann and Yu [13] reported that bagging is unnecessary for MARS (multivariate adaptive regression splines), but works for low-dimensional predictors, such as stumps. Furthermore they conclude that subbagging (a subsample aggregating) is as accurate as bagging, but computationally cheaper. More on the effectiveness of other resample schemes can be found in the studies of [15, 13, 14, 22] and can be understood in the perspective of creating diverse classifiers.

Bagging and its variants have been applied to enhance many learning techniques such as decision trees, Bayesian classifier or discriminat function [3, 5, 12]. Its application in rule induction algorithms is seldom [6, 8]. Stefanowski [6] reported that bagging substantially improved the predictive accuracy of the MODLEM rule induction algorithm. Our goal was to examine the application of bagging to the SA Tabu algorithm, also an algorithm inducing decision rules. We tested whether the bagged classifier trained by the SA Tabu Miner could achieve a better predictive accuracy than the single SA Tabu Miner algorithm. Moreover, as a promising variant of bagging we analyzed the subagging (subsample aggregating) method. We choose the subsample size m to be a fraction $m = a*N$ with $0 < a < 1$ and N is the size of the original training set. In addition, we also tried a modified version of the bootstrapping technique, where the repetition of instances was forced not to exceed two.

3 Bagging the SA Tabu Miner

In this section, we describe the application of the bagging method over the SA Tabu Miner algorithm. A detailed description of the bagged algorithm is shown in pseudo code in Figure 1. The training data is first resampled or subsampled and bootstrap samples are created. Then using each bootstrap sample, a classifier is created by the SA Tabu Miner algorithm. The final classifier is formed by performing a majority voting scheme over the generated classifiers.

```
DiscoveredClassifiersList = [ ]; /*initialized with an empty list*/
While (number of classifiers planned are not generated)
TrainingSet = {Determined % of all training cases};
DiscoveredRuleList = [ ]; /*initialized empty list*/
While (TrainingSet > Max_uncovered_cases)
Calculate entropy and hence probability
Start with an initial feasible solution S ∈ Ω.
Initialize temperature
While ((temperature > MinTemp) && (noDeadlock))
     Generate neighborhood solutions V* ∈ N(S).
     Update tabu timeouts of recently used terms
     Sort by (quality/tabu order) desc sol-s S* ∈ V*
     S* = the first solution ∈ V*
   While (move is not accepted or V* is exhausted)
        If metrop(Quality(S) - Quality(S*)) then
           Accept move and update best solution.
           Update tabu timeout of the used term
            break while
        End if
        S* = next solution ∈ V*
     End while
     Decrease temperature
End While
Prune rule S
Add discovered rule S in DiscoveredRuleList
TrainingSet = TrainingSet - {cases covered by S};
End While
Calculate DiscoveredRuleList predictive ability
Add DiscoveredRuleList in DiscoveredClassifiersList
End While
```

Where
(i) Ω is set of possible solutions,
(ii) S is the current solution and S* is the best found solution yet,
(iii) Quality(S) is cost function, which values the quality of the rule,
(iv) N(S) is neighbor of the solution S, and V* is sample of the neighbor solutions.

Fig. 1. Bagged SA Tabu algorithm

At the beginning, the list of discovered rules in every classifier is empty and the training set consists of all the training cases. Each iteration of the outer WHILE loop of SA Tabu Miner, corresponding to a number of executions of the inner WHILE loop, discovers one classification rule for one classifier. After the rule is completed, it is pruned from the excessive terms, to exclude terms that were wrongfully added in the construction process. The created rule is added to the list of discovered rules, and the training cases covered by this rule are removed from the training set. This process is iteratively performed while the following condition is satisfied

```
number of uncovered train cases > Max_uncovered_cases
```

where *Max_uncovered_cases* is a user-specified number of the unclassified training samples, usually 5% of all cases.

The selection of the term to be added to the current partial rule depends on both a problem-dependent heuristic function (entropy based probability), a tabu timeout for the recently used attribute values and the metropolitan probability function based on the Boltzman distribution of probability. The algorithm keeps adding one term at a time to its partial rule until one of the following two stopping criteria is met:

- Any term to be added to the rule would make the rule cover a number of cases smaller than a user specified threshold, called *Min_cases_per_rule*.
- The control parameter “temperature” has reached its lowest value.

Every time an attribute value is used in a term added to the rule, its tabu timeout is reset to the number of values of the particular attribute. In the same time, all other tabu timeouts of the other values for the particular attribute are decreased. This is done to enforce the use of various values rather than the most probable one, since often the difference in probability between the most probable one and the others is insignificant. Therefore, the final solution might not include the most probable values in the rule terms, but a combination of less probable ones.

In some cases, when two or more attribute values have close probabilities which are significantly greater than the probabilities of the other attributes, a dead lock race starts among the algorithms. In the next iterations all attributes with close probabilities are switching places and this continues until the minimal SA temperature is reached. In order to avoid this behavior, there is a tabu list for preventing dead lock cycles, which contains recently appeared qualities of proposed solutions and their appearing frequency. If a solution from the tabu list appears given number of times (*DeadLockTimes*), then the iteration cycle ends, returning the best found rule.

The entropy based probability guides and intensifies the search into promising areas (attribute values that have more significance in the classification), therefore intensifying the search. The tabu timeouts of the recently used attribute values discourage their repeated use, therefore diversifying the search. The metropolitan function controls the search, allowing greater diversification and a broader search at the beginning, while the control parameter temperature is big, and later in the process, when the temperature is low, it intensifies the search only in promising regions.

According to the metropolitan probability, a simulated thermodynamic system changes its state from energy E_1 to E_2 with probability:

$$P_T(\Delta E(X)) = \begin{cases} 1, \Delta E(X) <= 0 \\ \exp(-\dfrac{\Delta E(X)}{T}), \text{otherwise} \end{cases}$$

Where $\Delta E = E_2\text{-}E_1$, and T is the control temperature parameter.

4 Experimental Results

In this paper we examined the application of bagging over the SA Tabu Miner [1] algorithm, which is a rule induction algorithm, and checked whether the bagged classifier trained by the SA Tabu Miner could achieve better predictive accuracy than the single SA Tabu Miner algorithm. Moreover, we wanted to identify conditions under which the bagging method may improve the final prediction accuracy. Thus, we extended the bagging method with different resampling techniques:

- the resampling without replacement technique, where the subsample size m is $m = a*N$ with $0 < a < 1$ and N i is the size of the original training set and,
- the resampling with replication (bootsraping) technique, but the repetition of instances is forced to be not more than two.

The predictive accuracy is evaluated using the ten-fold cross validation technique [21]. The whole dataset is partitioned to 10 equal-sized blocks with similar class distributions, which are then grouped in a 9 to 1 ratio blocks. Each block is in turn used as a test set, while the classifier is trained over the remaining nine blocks. The whole process is executed 10 times. The final result is then averaged. An ensemble of T sub-classifiers is created over each fold. All subclassifiers (single classifiers inside an ensemble) are induced by the SA Tabu Miner algorithm. The maximal temperature in the algorithm is set:

$$MaxTemperature = InitialNumberOfSamples * NumberOfAttributes$$

The aforementioned value has been shown to give best results in single usage of our algorithm in comparison with the other rule induction algorithms.

All the experiments were performed using seven public-domain datasets from the UCI (University of California at Irvine) machine learning repository [20]. A brief overview of the datasets features is given in Table 1.

Table 1. Brief description of datasets

Data Set	Number of examples (instances)	Number of attributes
Hepatitis	155	19
Haberman's Survival Data Set	306	3
Bupa liver disorders	345	7
New thyroid gland data	215	5
Wisconsin breast cancer	569	32
Ljubljana breast cancer	286	10
Echocardiogram	132	12

When creating the bootstrap samples we used resampling without replacement with different resample sizes; we tried *m* being {95%, 90%, 75%} of *n*, the original training set size.

The number of bootstrap samples *T* is an important parameter, not only for the predictive accuracy, but for computational considerations, as well. Quinlan obtained good results for small numbers such as 3, 7 and 10 [12], but Brieman tested bagging using 10 to 100 bootstrap samples and Fei Xia [17] reported best results using only 10 bootstraps. Therefore, we choose *T*, the number of classifiers inside the ensemble, to be 10 and 50. In addition, we also tried a modified version of the resampling with replacement (bootstrapping) technique, where the repetition of instances would not exceed two.

The results are given in Table 2. For each data set, the first column shows the predictive accuracy obtained by a single classifier, while the next two columns contain the results of the basic version of the composite bagged classifier for different numbers of bootstrap samples. The followinggroup of columns, holds the predictive accuracy of the bagging method where one instance may appear twice at most.

Table 2. Results of different bagging approaches

Data set	Single SA Tabu Miner	Bagging with different *T*		Bagging where repetition does not exceed two	
		$T = 10$	$T = 50$	$T = 10$	$T = 50$
Hepatitis	89.2	76	81.33	82.22	84
Haberman's Survival	74.86	74.49	74.65	72.33	73.5
Bupa liver disorders	67.7	62.65	62.06	66.67	58.82
New thyroid gland data	92.44	87.14	88.57	90.79	89.76
Breast cancer Wisconsin	90.3	87.16	89.4	89.55	89.78
Ljubljana breast cancer	65.1	74.29	71.79	72.14	72.32
Echocardiogram	54.4	51.67	52.5	52.08	50.83

It is evident from the results that the bagging method applied to SA Tabu Miner achieves certain improvements to the predictive accuracy. Best improvement is obtained for the Ljubljana breast cancer data set. However, there are cases where due to the small number of examples in the training set, bagging gives no improvement at all. For the Hepatitis and New thyroid data it even decreases the predictive accuracy.

Comparing the results obtained using plain resampling with repetition and resampling with repetition where the repetition of instances does not exceed two, we conclude that better accuracy is obtained when the instance repetition is limited.

In order to further analyze the effect of the bagging ensemble method, we conducted experiments with different training data sizes. When creating the bootstrap samples we used resampling without replacement with different resample sizes. We tried *m* being {95%, 90%, 75%} of *N*, the original training set size. Buja and Stuetzle have shown that bagging is beneficial for $M_{\text{with\_replacement}} > N/6$ and $M_{w/o} > N/7$, but optimal is $M_{\text{with\_replacement}} = N/3$ and $M_{w/o} = N/4$. The results are shown in Table 3.

Table 3. Bagging without repetition with different T and different m

Data set	$T = 10$ $m = 75\%$	$T = 50$ $m = 75\%$	$T = 10$ $m = 90\%$	$T = 50$ m = 90%	$T = 10$ m=95%
Hepatitis	82.67	82.67	81.33	79.33	82.67
Haberman's Survival	72.67	73.67	73	73	72.33
Bupa liver disorders	57.65	61.18	66.18	64.12	62.35
New thyroid gland data	91.67	90.48	90	91.9	90.48
Breast cancer Wisconsin	90.75	88.36	91.49	91.19	90.45
Ljubljana breast cancer	72.86	71.43	72.5	72.14	72.5
Echocardiogram	53.33	51.67	50.83	52.5	50.83

The results of the experiments in Table 3 illustrate that best achievements are gained when using 90% of the base training set, while the number of bootstrap samples, that is voting classifiers, is only 10.

By comparing the results in Table 2 and Table 3, it can be concluded that bagging and its extensions does not affect the predictive accuracy significantly in every data set. Additionally, due to the increased number of voting classifiers, there is a loss in the comprehensible, simple and interpretable structure of the generated rules. In the Table 4, for each dataset, the average number of rules is listed. As expected, the number of rules increases as the number of classifiers rise.

Table 4. Average number of rules for each dataset

Data set	Hepatitis	Haberman Survival	Bupa	New thyroid	Breast cancer Wiscon.	Ljublj. breast cancer	Echocard.
No bag	3.21	6.4	9.9	5.8	6.1	8.55	8.3
$T = 10$	7.92	7.53	12.91	10.35	12.75	7.2	9.05
$T = 50$	9.09	9.31	13.92	12.15	15.37	8.71	10.5
$T = 10$	8.59	7.68	11.74	9.74	11.87	7	9.26
$T = 50$	9.38	8.66	13.13	11.71	14.07	7.69	10.76
$T = 10$ m=75%	8.3	7.62	11.54	9.58	12.24	6.91	9.7
$T = 50$ $m = 75\%$	9.72	8.881	14.23	11.69	14.34	7.8	10.45
$T = 10$ $m = 90\%$	8.41	7.44	12.31	9.41	11.54	6.62	9.22
$T = 50$ $m = 90\%$	9.29	8.82	13.87	11.64	14.04	7.73	10.99
$T = 10$ $m = 95\%$	8.55	7.16	11.64	9.14	11.95	6.56	9.79

5 Conslusion

This paper presented the implementation of bagging techniques to the SA Tabu Miner algorithm for classification rule induction. We have compared the performance of various version of bagging techniques applied to SA Tabu Miner on public domain data sets. The results showed that certain bagging approaches with particular parameters can significantly increase the predictive accuracy of the algorithm. However, in some fewer datasets, due to their statistical nature and the number of instances, bagging does not bring any benefits. Thanks to the resampling component, bagging can be useful in poor training sets, or when it is impossible to get multiple samples. On the other hand, it cannot be applied in cases where there are small data sets, or the original sample is not a good approximation of the population.

Since comprehensibility is important whenever discovered knowledge will be used for supporting a decision made by a human user, SA Tabu Miner often discovered simpler rule lists. However, with the application of bagging the advantage of comprehensibility is lost for the sake of greater predictive ability. In scenarios where human verification of the discovered knowledge is not important and maximized predictive accuracy is the goal, the approaches proposed in the paper can be applied.

References

[1] Chorbev, I., Mihajlov, D., Jolevski, I.: Web Based Medical Expert System with a Self Training Heuristic Rule Induction Algorithm. In: Proc. of The First International Conference on Advances in Databases, Knowledge, and Data Applications, DBKDA 2009, Cancun, Mexico, March 2009, pp. 143–148 (2009)

[2] Dietterich, T.G.: Machine Learning Research: Four Current Directions. AI Magazine 18(4), 97–136 (1997)

[3] Dietterich, T.G.: Ensemble Methods in Machine Learning, Oregon State University, Corvallis, Oregon, USA, tgd@cs.orst.edu. WWW home page, `http://www.cs.orst.edu/tgd`

[4] Gentle, J.E., Härdle, W., Mori, Y.: Handbook of Computational Statistics, ch. 16. Springer, Heidelberg, `http://fedc.wiwi.hu.berlin.de/xplore/ebooks/html/csa`

[5] Breiman, L.: Bagging predictors. Machine Learning 24(2), 123–140 (1996)

[6] Stefanowski, J.: Bagging and Introduction of Decision Rules. In: Klopotek, M., et al (eds.) Intelligent information systems (2002)

[7] Freund, Y., Schapire, R.: Experiments with a new boosting algorithm. In: Proc. of the Thirteenth Int. Conf. on Machine Learning, pp. 148–156 (1996)

[8] Pino-Mejías, R., et al.: Bagging Classification Models with Reduced Bootstrap. Structural, Syntactic, and Statistical Pattern Recognition, 966–973 (2004), `http://www.springerlink.com/content/6r0b2payc24fj93e/`

[9] Grandvalet, Y.: Bagging equalizes influence. Machine Learning 55(3), 251–270 (2004)

[10] Friedman, J.: Stochastic gradient boosting. Computational Statistics and Data Analysis 38(4), 367–378 (2002)

[11] Parmanto, B., Munro, P., Doyle, H.: Improving Committee Diagnosis with Resampling Techniques. In: Touretzky, D., Mozer, M., Hasselmo, M. (eds.) Advances in Neural Information Processing Systems, vol. 8, pp. 882–888 (1996)

[12] Quinlan, J.R.: Bagging, boosting and C4.5. In: Proceedings of the 13th National Conference on Artifitial Intelligence, pp. 725–730 (1996)

[13] Bühlmann, P., Yu, B.: Explaining Bagging'. Technical Report 92, Seminar für Statistik, ETH, Zürich (2000)

[14] Buja, A., Stuetzle, W.: The Effect of Bagging on Variance, Bias and Mean Squared Error. Technical report, AT&T Labs-Research (2000)

[15] Friedman, J.H., Hall, P.: On Bagging and Non-linear Estimation. Technical report, Stanford University, Stanford, CA (2000)

[16] Quinlan, J.R.: C4.5: Programs for Machine Learning. Morgan Kaufmann, San Francisco (1993)

[17] Xia, F.: Bagging (2006), http://faculty.washington.edu/fxia/courses/LING572/pagging.ppt

[18] Buja, A., Stuetzle, W.: Observation of Bagging. Statistica Sinica 16, 323–351 (2006)

[19] Hall, P., Samworth, R.J.: Properties of Bagged Nearest-neighbor Classifiers. J. Roy. Statist. Soc., Ser. B 67, 363–379 (2005)

[20] http://archive.ics.uci.edu/ml/

[21] Weiss, S.M.: Small Sample Error Rate Estimation for k-neares Neighbor Classifiers. IEEE Transaction of pattern analysis and Machine Intelligent 13(3), 285–289 (1991)

[22] Brown, G., Wyatt, J., Harris, R., Yao, X.: Diversity Creation Methods: A Survey and Categorisation. Information Fusion 6(1), 5–20 (2005)

[23] Hansen, L., Salamon, P.: Neural network ensembles. IEEE Transactions on Pattern Analysis and Machine Intelligence 12, 993–1001 (1990)

[24] Krogh, A., Vedelsby, J.: Neural network ensembles, cross validation, and active learning. In: Tesauro, G., Touretzky, D., Leen, T. (eds.) Advances in Neural Information Processing Systems, vol. 7, pp. 231–238. MIT Press, Cambridge (1995)

[25] Hashem, S.: Optimal linear combinations of neural networks. Neural Networks 10(4), 599–614 (1997)

[26] Opitz, D., Shavlik, J.: Actively searching for an effective neural-network ensemble. Connection Science 8(3/4), 337–353 (1996a)

[27] Opitz, D., Shavlik, J.: Generating accurate and diverse members of a neural-network ensemble. In: Touretsky, D., Mozer, M., Hasselmo, M. (eds.) Advances in Neural Information Processing Systems, vol. 8, pp. 535–541. MIT Press, Cambridge (1996b)

Information/Material Processing Synergy: Flexible Manufacturing and Operating System Metaphor for a Biological Cell

Nevena Ackovska[1] and Stevo Bozinovski[2]

[1] Institute of Informatics, University St. Cyril and Methodius,
1000 Skopje, Macedonia
nevena@ii.edu.mk
[2] South Carolina State University,
Orangeburg, SC, USA
sbozinovski@scsu.edu

Abstract. This paper presents the information/material processing synergy in both biological and human made systems. It is a further elaboration on the metaphors previously proposed for genetic information processing, such as the robotics/flexible manufacturing metaphor and the cell systems software metaphor. Related issues are also discussed, such as file system, program preparation and its parallel and distributed features, including interthread communication, among others. The paper proposes that, from a manufacturing science viewpoint, the protein biosynthesis process can be viewed as a CAD/CAM system for molecular biology.

Keywords: manufacturing science, information/material processing synergy, cell operating system, cell CAD/CAM system, distributed systems, protein biosynthesis multithreading, nanotechnology.

1 Introduction

The understanding of genetic information processing evolved through several metaphors. A metaphor represents paradigm transformation – we use knowledge from a familiar system in order to understand and develop solutions in another system. The first metaphor explaining genetic processes was the biochemistry metaphor, which basically relied on the fact that “DNA is an acid”. The important breakthrough was the linguistic and information processing metaphor which stated that “DNA is a string of letters” [1], [2]. We proposed the robotics and flexible manufacturing metaphor [3], which pointed out that “DNA contains programs for machines and robots, and for example tRNA is a mobile robot”. The next metaphor we proposed [4] is the operating systems and systems software metaphor, which states that “DNA is the cell real-time distributed operating system”. This paper continues our research in understanding genetic information processing using the metaphors we proposed.

The idea is that a biological cell has some kind of distributed operating system, which resides on several minidisks (chromosomes) [4], [5]. The paper we are presenting here elaborates further in that direction, emphasizing that in addition to being a *parallel distributed information processing* system, a genetic system is also a real-time and *parallel distributed material processing* system [3], [6]. The basis for the material/information processing synergy is presented in Table 1.

Table 1. Information/material synergy

	Information	Material
Processing	Information processors	Material processors
Transport	Information transmission	Material transportation
Storage	Information bases	Material bases

The information/material synergy is interesting for both computer science and manufacturing science. In the following sections, we will focus the attention to the parallel and distributed information/material processing system of a biological cell. In this paper we also propose that a CAD/CAM concept is being used in genetic manufacturing.

2 The Cell and Its Parallel and Distributed File Processing Features

When we, as computer engineers, think of parallel features of computer systems, we always imagine segments of code that could be processed in parallel. Sometimes, we have the opportunity to design the hardware components of the system in order to maximize the parallel features of the software we work with. Sometimes we are able to design both, the hardware and the software for the needed system, and make its parallel features as optimal as the price/performance ratio allows us.

With the living beings this conformism is not applicable. When there is a specific situation, the living system has to act in a certain, fast manner, so it could survive. Therefore, the parallel and distributed features of the living systems are incorporated in the way they are constructed.

Here we will present some analogies between human made systems and biological systems regarding the parallel and distributed nature of information and material processing systems.

2.1 Cell Files

When studying genetics, the crucial concept is the concept of a gene. Thus, a very natural question is “What is a gene?” A usual answer is that a gene is a segment of DNA that encodes for either protein or RNA. Also, one could encounter slightly different definitions [7].

In computer science and engineering, a usual reasoning about an information processing system considers the files of that system. So, for genetic information processing we might ask the question "*what are the files of the genetic information processing system?*" Is the concept of a gene analogous to the concept of a file? Having this as a starting point, in this section we will present our understanding of DNA organization and DNA computing in terms of files and related concepts.

Looking for a concept of a file in DNA, we found that the transcription units (or scriptons [8]) are analogous to cell files. A transcription unit is a segment of DNA that eventually becomes transcribed to RNA. In prokaryotes (cells without a nucleus), a transcription unit often produces a transcript with several genes (so-called polycistronic RNA). In eukaryotes (cells with a nucleus containing DNA) it produces a precursor RNA (pre-RNA), which contains the information about a single gene, but in order to obtain it, additional processing needs to be performed.

The eukaryotic files are rather complex and contain segments of a gene, interleaved with segments that do not belong to the gene. Those segments are known as introns (interleaving segments), as opposite to exons (gene expressing segments). To the people involved with genetics, there is a standard question considering this phenomenon: how did it happen that eukaryotic genes became segmented? However, for computer engineers introduced to the concept of a file, the answer is straightforward – busy files are fragmented. Defragmentation is sometimes needed in computer file systems. Moreover, it is expected that between two fragments of a file an entire different file could be placed. This fact points to the concept of distributed file systems [9], [10]. And indeed this is the case in molecular genetics: after the first evidence that Tetrahymena ribozyme is actually an intron [11], more evidence has been found that genetic files could be found within a different file [12]. Therefore, our file-centered approach offers simple answers to nontrivial problems in genetics [13].

Now, let us consider the exons, the gene expressing segments. Using our approach, it is easy to see that exons could be functional units, such as subroutines (or methods in OOP) of a more complex program file. The subroutines could be reusable, meaning that the same exon could be used in different RNA's. An example of such a phenomenon could be the building of mRNAs for antibody proteins.

We believe that while the genes are the proper concept when talking about heredity, we propose that the concept of a file is very useful in describing the DNA transcription process. This makes the first step in the analogy between computer systems and genetic systems. We believe that the whole transcription process could be better explained in file processing terms instead of linguistic terms. The cell, especially the eukaryotic cell, undergoes extensive file processing: from copying the pre-RNA file to obtaining the RNA message. This process includes operations like: cut (introns), join (exons), right append (trailer string), left append (header string), letter replacement and so on, which are standard file processing operations in every modern computer operating system [4].

2.2 Cell Disks

Following the same line of reasoning, another very important question would be: what is a disk in a genetic information system? The initial observation could lead to the conclusion that the DNA offers a tape based [14] information processing system. However, a careful examination reveals that DNA tapes are randomly accessible – they are not sequential tapes. So, due to the nature of their accessibility, we may consider them as disks. It is important to observe that DNA can easily be represented by the classical concept of a cylinder. The cylinder means that, although it is on different disk plates, the information can be read in the same time by several disk reading heads. In a sense, the concept of a cylinder enables spatially distant information to be processed in parallel. And indeed, the cell disk (the complete genome) is distributed over several disk plates, or minidisks (chromosomes). The files on different chromosomes could be processed in parallel. For example, the files needed for producing an important organism substance, the hemoglobin, are distributed over both chromosome 11 and chromosome 16; parallel distributed processing is needed in order to obtain a single molecule of hemoglobin.

2.3 Process Management vs. Memory Management

Classical process management in computer systems considers a resource and several processes competing over that resource. Several mechanisms have already been developed to address this issue [15]. Basically, a mutual exclusion mechanism is used, where one process is using a resource while other processes are blocked, waiting.

It seems that in genetic systems that mechanism is not employed. Instead of process management, genetic systems rather utilize memory management. They simply keep many copies of files that are frequently used. It seems that the redundancy in the genetic system really enables parallelism. It also enables more space where the processes could be processed, since the files themselves are spatially distant. The execution can be run on several processors, distributed spatially.

Examples of such files are rRNA files needed for production of ribosomes. Those files are kept in thousands of copies [7]. In order to function in real time, a genetic system needs to produce many copies of the rRNA in the same time, in order to respond to a change in the environment. This means that no waiting for this vital resource is allowed. So, in genetic systems, in order to achieve a real-time response, many copies of rRNA files are utilized, and they are all processed in parallel, and distributed over a vast region in the cell. They are produced to become part of ribosomes, the cell processors.

During evolution, the files could be produced in as many copies as needed. This is a possible cause for file fragmentations. It also gains to redundancy. But, each process could work on its own copy of the same file. This is a space management mechanism, rather than a time management one. It is interesting that the novel approaches toward operating systems design propose a similar prospective. In the Nooks approach [16] the kernel module is cloned and then the clone is modified outside the kernel. The cloning mechanism is used in resolving the issues with threads, where instead of blocking a thread on access to shared data, a proxy of the data is generated, like in the Promisses approach [17].

There is another, very popular and rapidly developing system, which supports the claim that space management tends to replace process management. The Google File System (GFS) [18] is a scalable distributed file system for large distributed data-intensive applications. It provides fault tolerance while running on inexpensive commodity hardware, and is used by a large number of users. The large cluster used provides hundreds of terabytes of storage across thousands of disks on over a thousand machines, and it is concurrently accessed by hundreds of clients. The design space for this file system emphasizes that the space management, file redundancy and statistically based algorithmic approach are the key concepts in making GFS the supreme contemporary file system.

2.4 Time-Sharing vs. Autonomy

It seems that the time-sharing concept is not applied in the living systems. Their components are built to be extremely autonomous while they have a task to run. If the process is vital, no waiting for a processor is allowed. Instead, the new routine, run by a newly built processor will be started, it will build the needed autonomous processor, distribute it through the cell, and the desired process will be started. And, most probably, many copies of that same process will be started on many newly built processors for solving that particular problem.

2.5 Consistency

Keeping files in many copies requires a mechanism of file consistency. It seems that the voting mechanism is used in a cell to keep some files unchanged in the evolution process. For example, the rRNA files are particularly stable. Also, the fact that the DNA is a two-strand molecule, in which one strand is a complement to the other, enables repair of an error and consistency in obtaining copied information from a DNA strand [7].

2.6 Pipeline Processing

The cell system is a highly distributed one and many processes are running in parallel and some of them are running in a pipeline. One prominent example is the polysome, sequence of several ribosomes over a single mRNA in prokaryotes. Many ribosomes are actually accessing the same file, but in a pipeline manner!

2.7 Parallel Processing: Inter-thread Communication

In such a system it is possible to assume an existence of communicating threads that compose one process. In order to study possible mechanisms of such an inter-process communication, we developed a software model of a minimal protein biosynthesis system (Fig. 1).

Three agents – RNA polymerase, aminoacil-tRNA synthetase, and ribosome – are engaged in inter-thread communication during the protein biosynthesis. The process is initiated on demand for a protein which activates an RNA polymerase agent. It reads the DNA bases and appends an RNA base to the evolving RNA. Once an mRNA is released, a ribosome reads it, codon by codon. For each codon it looks for corresponding anticodon-tRNA^{aa}, which will bring the needed aminoacid *aa* (Fig. 1). Actually, the physical representations of the genetic code are the tRNAs. The ribosome assembles the protein and releases it. Another agent, aminoacil-tRNA synthetase, works in parallel with the ribosome and RNA polymerase. It assures that the tRNA is loaded with the corresponding aminoacid. It receives an appropriate amino acid, receives an appropriate tRNA, loads the amino acid to the tRNA, and then releases the loaded tRNA.

```
thread RNA polymerase
repeat
  receive demand (protein)
  initiate reading gene(protein)
  repeat
    read DNA_base from gene(protein)
    append mRNA_base(DNA_base) to mRNA
  until stopcondition(gene(protein))
  release  mRNA
until degradation
```

```
thread  $a_i$-tRNA synthetase
repeat
        receive $a_i$
        receive $tRNA^i$
        load  $a_i$ to $tRNA^i$
        release $a_i tRNA^i$(anticodon)
until degradation
```

```
thread Ribosome
begin
        receive mRNA
        repeat
           read codon from mRNA
           receive $a_i tRNA^i$ (anticodon(codon))
                    unload $a_i$ from $a_i tRNA^i$
                    assemble $a_i$  to protein
        until stopcondition(mRNA)
        release protein
recycle
```

Fig. 1. Understanding protein biosynthesis using inter-thread communication metaphor

Our model uses the **receive** and **release** primitives for inter-thread communication. The amino-acid handling part is modeled using robotic language commands, such as load, unload and assemble. Instead of using the primitives receive and release, one can use the standard primitives **receive** and **send**, or some classical primitives such as **wait** and **post** [19] which were used in some pioneering real-time multithreading robot control efforts [20], [21].

The software model above is not intended to be a detailed simulation of the protein biosynthesis process. Details involving proteins as initiation, elongation and termination factors are neglected. The model studies only possible inter-thread communication between the main protein synthesis agents. Petri nets have been also used for modeling the process shown above [4], [22].

2.8 Compilation as Manufacturing – From Source Code to a Robot

In building autonomous robots the software is developed separately from the hardware. A source code is compiled into an executable code and then an EPROM is produced that is embedded into a robot. Our research suggests that the robot producing process in cells is an extension of the compilation process: the cell operating system takes a source file (gene) and produces a robot in the form of an RNA (ribozyme) or a protein (enzyme). The cell robots and other cell machines are produced real-time, on demand. So, once demand is sensed by the DNA, it makes an appropriate source file accessible, out of which the requested program is read (the linear form of RNA). This program is then compiled and assembled into a robot by obtaining its 3D functional structure. An example of such a robot is the tRNA, which is a shuttle robot that carries amino acids into a ribosome for protein assembly. Another example is the rRNA, which, being a part of a ribosome, performs a protein assembly function. In addition to RNA robots, the cell produces protein robots, i.e. enzymes, which are the major working force in a cell. This includes carrier mobile robots like kinesin, which travels along the cell cytoskeleton, or a two-armed robot, lac-repressor in E. Coli, which grabs the DNA and disables its access.

Here we propose that protein manufacturing is actually the CAD/CAM (Computed Aided Design/Manufacturing) system in a cell. The CAD part consists of the genes carrying software for protein or RNA production, while the CAM part is the entire machinery (the genosome) that obtains the product described in a gene.

3 Just-in-Time Manufacturing

The cells are just-in-time (JIT) systems, rather than just-in-case (JIC) systems. When need is sensed, they produce the needed elements very promptly. Most of those processes are actually running in parallel: from parallel information processing, to parallel material processing in distributed sites in the cell. Not only the information, but all the products, including needed machinery and robots, are produced on demand. That makes them the ultimate real-time systems: from processing information in real time, through producing tools and robots in real time, to producing the final product.

In order to do so, the cell keeps many copies of the most needed programs, in order to access them all in parallel when need is sensed. Furthermore, it will produce all the needed machinery for processing these programs in parallel.

3.1 Common Building/Coding Blocks

Additional adaptation towards parallel processing of a cell is that it keeps just a small number of building blocks: there are only 20 different aminoacids used for protein synthesis. This enables rapid switching between demanded products and flexible manufacturing.

It should also be noted that the cell produces its software (e.g. mRNA) and its hardware (e.g. enzyme or ribozyme) in a similar way, by appending building blocks to a sequence which in turn folds into a 3D structure. It seems that this approach in interleaving the software and hardware should be the next step in human made systems.

4 Nanotechnology and the Future of Parallel Processing Systems

Parallel distributed computation is about speeding up the applications. Real-time application in general is due to the need of real-time product. Our study shows that the future parallel distributed processing systems needs to *integrate parallel distributed material and information processing*. And this is what the nature has done long ago.

Today scientists make efforts to bring information and material processing closer to one another. Nanotechnology science makes it possible to produce nanostructures that do some kind of processing while having the means to produce other nanorobots that do specialized work in cells (and possibly self-replication) [23]. These robots have their software and hardware interleaved in the same building-coding blocks. There is also a concern of using the combination of DNA information-encoding and recognition properties, and the enzymatic machinery capability for DNA manipulation in the field of DNA computation [24].

One open problem today is how to construct carrier nanorobots, for example robots that could transfer an atom, or a whole another molecule, from one place in the medium, to another place [25]. However, in 2005 there were some reports on nanorobots that could move in a desired direction, and were symbolically named nanocars [26], [27]. The future of these nanovehicles is to be able to carry controlled load, so they could be used for precise drug delivery. There has been a suggestion [28] that nanomachines based on the RNA are needed. It relates nanotechnology vehicles research with our early suggestion that tRNA is a mobile robot.

Having the aforementioned in mind, and the possibilities that nanotechnology science offers, we could presume that, perhaps just like the evolution of natural life, the evolution of artificial life will also continue from the nano level.

Of course, it is important that along the way we be interested in faster algorithms, for example for floating point division, but ultimately the effects of that algorithm will be used in some real-time application, for example robot jumping, or for just-in-time production. Therefore, probably it is sensible to try to speed up the whole process, just as the nature has done – create the hardware and the software of the machines from the same "material".

5 Conclusion

The integration of information processing and material processing is present in biological systems since the existence of life. Humans are now building flexible manufacturing systems realizing that they were actually invented by biology. Therefore, the information/material processing synergy, the concept we introduce in this paper, is essential for life.

Our research is based on two approaches: a biocybernetics one and a bionics one. The biocybernetics approach observes the biological cell as a flexible manufacturing system, controlled by a cell operating system and systems software. The bionic approach suggests improvement in human made systems by implementing the knowledge we've gained from studying biological systems.

Future systems should be built from modules that inherently contain features of mutual communication and self assembly, similar to proteins. The lesson that we could learn from molecular biology is that the compilation process in cells is actually a CAD/CAM process. We discussed that the cell robots and other OS embedded systems, such as lac-repressor or tRNA, are developed in a single program compilation/production line, not separately, as in today's technologies. To do so, the cell builds both hardware and software using the same material. These building blocks are constantly recycled, which is another important feature of genetic manufacturing and information processing.

The integration of material and information processing is already of great interest in the scientific community. Nanotechnology science makes it possible to produce nanostructures that do specialized work in cells. These robots have their software and hardware interleaved in the same building-coding blocks. Perhaps, just like the evolution of natural life, the evolution of artificial life will also continue with big steps from the nano level. In human-made systems this feature is not implemented, although there is some effort in this direction [29].

We believe that our approach based on flexible manufacturing and robotics offers better understanding of genetic processes. Conversely, from genetic information processing we have learned that the synergy between material and information processing is a crucial approach in the manufacturing and information processing and should be used in building human made systems.

References

1. Watson, J., Crick, F.: Molecular structure of nucleic acids: a structure of deoxyribose nucleic acid. Nature 171, 737–738 (1953)
2. Crick, F.: On protein synthesis. In: Proc. Symp. Society for Experimental Biology, vol. 12, pp. 138–163 (1958)
3. Bozinovski, S.: Flexible manufacturing systems: A biocybernetics approach. In: Vukobratovic, M., Popov, E. (eds.) Robotics and Flexible Manufacturing, Moscow, pp. 192–197 (1986)
4. Bozinovski, S., Jovancevski, G., Bozinovska, N.: DNA as a real time, database operating system. In: SCI 2001, Orlando, pp. 65–70 (2001)
5. Bolsover, S., Hyams, J., Jones, S., Shephard, E., White, H.: From Genes to Cells. Willey-Liss (1997)
6. Bozinovski, S., Bozinovska, L.: Manufacturing Science and Protein Biosynthesis. In: Callaos, N., Badawy, W., Bozinovski, S. (eds.) Systemics, Cybernetics, and Informatics, Orlando, pp. 59–64 (2001)
7. Brown, T.A.: Genomes, 2nd edn. Willey-Liss (2002)
8. Ratner, V.: Control Systems in Molecular Genetics, Nauka, Novosibirsk (1975)
9. Nutt, G.: Centralized and Distributed Operating Systems. Prentice-Hall, Englewood Cliffs (1992)
10. Tanenbaum, A.: Distributed Operating Systems. Prentice-Hall, Englewood Cliffs (1995)
11. Kruger, K., Grabowski, P.J., Zaug, A.J., Sands, J., Gottschling, D.E., Cech, T.R.: Self-splicing RNA: Autoexcision and autocyclization of the ribosomal RNA intervening sequence of tetrahymena. Cell 31, 147–157 (1982)
12. Been, M.: Versatility of Self-Cleaving Ribozymes. Science 313, 1745–1747 (2006)
13. Ackovska, N., Bozinovski, S., Jovancevski, G.: A New Frontier for Real – Time systems – Lessons from Molecular Biology. In: Proc. IEEE SoutheastCon, pp. 224–228 (2007)
14. Ackovska, N.: System software of minimal biological systems, PhD Thesis, Skopje (2008)
15. Andrews, G., Schneider, F.: Concepts and notations for concurrent programming. ACM Computing Surveys 15, 3–43 (1983)
16. Swift, M., Bershad, B., Levy, H.: Improving the reliability of commodity operating systems ACM Trans. Compuer Systems 23, 77–110 (2005)
17. Lee, E.: The problem with threads. Computer, 33–42 (2006)
18. Ghemawat, S., Gobioff, H., Leung, S.-T.: The Google File System. In: Proc. 19th ACM Symp. on Operating Systems Principles, Bolton Landing, NY, pp. 29–43 (2003)

19. Witt, B.: Communication modules: A software design model for concurrent distributed systems. IEEE Computer Magazine (1985)
20. Bozinovski, S., Sestakov, M.: Multitasking operating systems and their application in robot control. In: Proc. Workshop on Macedonian Informatics, Skopje, pp. 195–199 (1983) (In Macedonian)
21. Bozinovski, S.: Parallel programming for mobile robot control: Agent based approach. In: IEEE Conf. on Distributed Operating Systems, pp. 222–228. Computer Society Press (1994)
22. Balasubramanian, N., Yeh, M.-L., Chang, C.-T., Chen, S.-J.: Hierarchical Petri Nets for Modeling Metabolic Phenotype in Prokaryotes. Industrial and Engineering Chemistry Research 44, 2218–2240 (2005)
23. Liao, S., Seeman, N.C.: Translation of DNA Signals into Polymer Assembly Instructions. Science 306, 2072–2074 (2004)
24. Zhang, M., Tao, W., Tarn, T.-J., Xi, N., Li, G.: Interactive DNA Sequebce and Structure Design for DNA Nanoaplication. IEEE Transactions on Nanobioscience 3(4), 286–292 (2004)
25. Hogan, J.: DNA robot takes its first steps. Journal reference: Nano Letters, http://www.newscientist.com/article.ns?id=dn4958, doi:10.1021/n1049527q
26. Morin, J.-F., Shirai, Y., James, M.: En Route to a Motorized Nanocar. Organic Letters 8(8), 1713–1716 (2006)
27. Shirai, Y., Osgood, A.J., Zhao, Y., Kelly, K.F., Tour, J.M.: Directional Control in Thermally Driven Single-Molecule Nanocars. Nano Letters 5(1), 2330–2334 (2005)
28. Chworos, A., Severcan, I., Koyfman, A.Y., Weinkam, P., Oroudjev, E., Hansma, H.G., Jaeger, L.: Building Programmable Jigsaw Puzzles with RNA. Science 306, 2068–2072 (2004)
29. Demeester, L., Eichler, K., Loch, C.H.: Organic Production Systems: What the Biological Cell Can Teach Us About Manufacturing. Manufacturing and Service Operation Management 6(2), 115–132 (2004)

Position Estimation of Mobile Robots Using Unsupervised Learning Algorithms

Petre Lameski and Andrea Kulakov

Faculty of Electrical Engineering and Information Technologies, Skopje, R. Macedonia,
Karpos 2 bb, P.O. Box 574, Skopje, R. Macedonia
{lameski,kulak}@feit.ukim.edu.mk

Abstract. Estimating the position of a mobile robot in an environment is a crucial issue. It allows the robot to obtain more precisely the knowledge of its current state and to make the problem of generating command sequences for achieving a certain goal an easier task. The robot learns the environment using an unsupervised learning method and generates a percept – action- percept graph, based on the readings of an ultrasound sensor. The graph is then used in the process of position estimation by matching the current sensory reading category with an existing node category. Our approach allows the robot to generate a set of controls to reach a desired destination. For the learning of the environment, two unsupervised algorithms FuzzyART neural network and GNG network were used. The approach was tested for its ability to recognize previously learnt positions. Both algorithms that were used were compared for their precision.

Keywords: Mobile robots, Position estimation, Unsupervised learning.

1 Introduction

Position estimation is a challenging task in mobile robotics, especially when dealing with a previously unknown environment. The mobile robot must be able to learn the environment by memorizing the states that it has previously visited and in the same time recognize the previously visited positions. Position estimation is also used in the problems of environmental mapping and localization problems. There are many approaches in the literature for both environment learning and mapping problems and localization and position estimation problems. [1][2]

Many approaches for solving the mapping and localization problem are proposed in the literature. For the environmental learning problem, the occupancy grid map is an elegant way of representing the knowledge about the environment. For the problem of localization in a known environment some of the proposed approaches are Grid localization, Montecarlo localization and many other localization approaches [1]. These proposed approaches use the previous knowledge of the environment (a map of the environment) and localize the robot in the known environment.

The problem of Simultaneous Localization and Mapping (SLAM), which urges the robot to learn the environment and estimate its position in the same time, has many different approaches [1]. There are many proposed variations of SLAM algorithms based on Kalman filters, Particle filters and other types of filters that give good results in the SLAM problem and are widely used.

The action planning problem is also an important issue in mobile robotics. One of the most widely used approaches are the Markov decision process (MDP) and the partially observable MD. These methods, however, are computationally expensive for large maps [1].

Artificial neural networks have been used for the problem of environment learning and localization [2]. They give good results in qualitative localization but not in fine localization.

Artificial neural networks are also widely used in the area of object recognition, classification [3] and in control problems [4] where they are used for robot navigational behavior.

For a mobile robot to be able to automatically move and in the same time learn an unknown environment all of these problems must be solved in the same time, thus the need of combining the solutions in a single system. A fully autonomous robot must be able to perform the previously stated actions.

This approach is proposing a solution that combines the environmental learning, position estimation and action planning in a single solution, as it was presented in [5]. In this paper we are using GNG network for the learning of the environment, and comparing it with the FuzzyART network, used for the same purpose. The mobile robot is behaving autonomously and explores the environment by learning the new positions and in the same time recognizing the previously visited position. The obtained knowledge of the environment can be used for generating control sequence for the robot to be able to reach a desired position.

2 System Architecture

In this paper we are using an integrated system for environmental learning, position estimation and planning of actions of a mobile robot [5]. The goal of this system is for the robot to be able to learn a previously unknown environment and be able to use the knowledge about the environment to generate actions for achieving a certain goal or a certain position in the environment.

The system is using a Percept – Action – Percept model [6] [7] of learning, in which, the positions of the robot are evaluated by the sensor readings. Each position is connected to another position by the action that the robot performed or needs to perform in order to reach that position. The Percept – Action –Percept model is shown in Fig 1.

The data about the environment is organized as a connected graph. The nodes contain the information of the robot state, and the links contain the information about the action that is needed for transition from one state into another as in Fig 2. This type of data organization makes the problem of action planning very simple. All the robot needs to do is to use a graph search algorithm and find the action sequence needed to reach the desired position. During the process of executing the actions, the robot must verify that its current position is the one that is desired by the action plan.

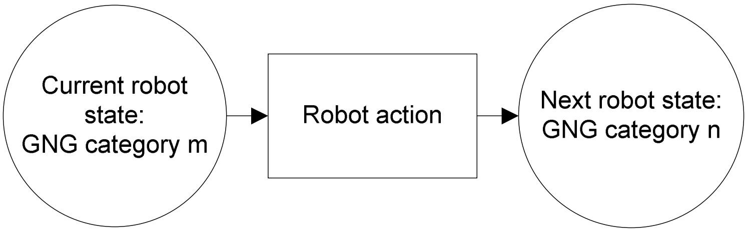

Current robot state: GNG category m
Robot action
Next robot state: GNG category n

Fig. 1. Percept – action – percept model

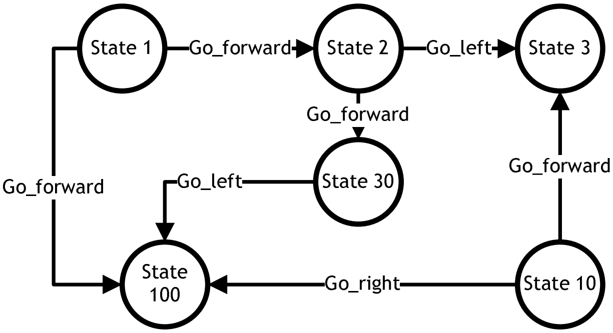


Fig. 2. State graph representation of the environment with typical actions

Each percept at the robot's position is categorized by using an unsupervised learning algorithm. Two algorithms, which are used for that task, are the FuzzyART neural network [8] [9] in the first approach (already considered in [5]) and the GNG network [10] [11] in the second approach (considered in this paper). A Lego Mindstorms NXT robot is used as a robot platform and its ultrasound sensor is used for perceiving the environment. The robot uses two servo motors for the movement in the environment and one servo motor to rotate the ultra-sound sensor in different directions. For each robot position, 12 measurements in different directions are taken. These measurements define the percept of the position.

The GNG network is used for remembering and estimating the positions of the mobile robot in the environment. Each percept acquired by the robot is normalized and then used as input in the GNG neural network. After a category is selected, the graph is checked for an existing node. If none exists, the new category is added to the graph as a new state and a new action link is created from the previous category. If the category exists, then if necessary, a new action link, containing the necessary action as in Fig. 2, is created to connect the previous one with this category. The state represents the position in which the robot is at the moment. Different state means that the robot estimated it is as a different position. The percept in a robot position is used for determining the state. In Fig. 3 is given the activity diagram of the developed processes.

Adaptive resonance theory (ART) networks develop stable recognition codes by self-organization in response to arbitrary sequences of input patterns. They are able to continue to learn from new events without forgetting previously learned information. ART networks model several features such as robustness to variations in intensity, detection of signals mixed with noise, and both short- and long-term memory to accommodate variable rates of change in the environment. There are several variations of ART-based networks, but we have used FuzzyART, which has analog inputs, as most appropriate for these kinds of tasks [5].

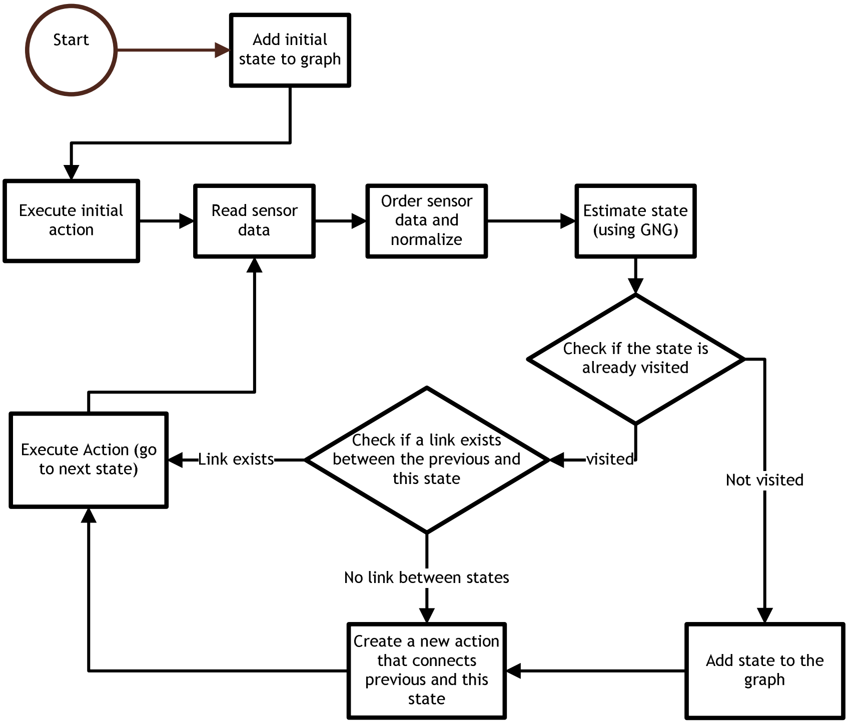

Start
Add initial state to graph
Execute initial action
Read sensor data
Order sensor data and normalize
Estimate state (using GNG)
Check if the state is already visited
visited
Not visited
Check if a link exists between the previous and this state
Link exists
Execute Action (go to next state)
No link between states
Create a new action that connects previous and this state
Add state to the graph

Fig. 3. UML Activity diagram

The GNG algorithm is an unsupervised incremental clustering algorithm. Given some input distribution in the input space R^n, GNG incrementally creates a graph, or network of nodes, where each node in the graph has a position in R^n. GNG is an adaptive algorithm in the sense that if the input distribution slowly changes over time, GNG is able to adapt, that is to move the internal nodes, so as to cover the new distribution. The algorithm constructs a graph in which nodes are considered neighbours if they are connected by an edge. The neighbour information is maintained throughout the execution by a variant of competitive Hebbian learning [11]. The nodes of the GNG graph tend to follow the distribution of the signals. Each node in the GNG algorithm is consisted of the following:

- a reference vector, in R^n,
- a local accumulated error variable, and
- a set of edges defining the topological neighbours of the node k.

The reference vector represents the node in the signal space. The local accumulated error is a statistical parameter that is used for determination of the place where a new node should be inserted. Each edge between the nodes has an aging parameter used for removing old and inactive edges. This is necessary since the nodes are moved.

The GNG algorithm first initializes two nodes, and links them. For each signal that is received, a winner node (the node that is closest to the signal in R^n space) is found and the second closest node is found. The winner node is moved towards the signal and all topological neighbours are moved towards the signal too. The error variable is updated for the winner node and its topological members. Then the age of the edges of the winner node are reset to 0. If there is an edge between the winner node and the second closest node, it is reset to 0, if an edge does not exist, it is created. If a creation criterion is met, usually when the number of iterations reaches a certain multiple of a constant number λ, a new node is created between the nodes that have the largest accumulated error. The error of these nodes is updated. Edges that reach a certain pre-specified age and nodes that don't have topological neighbours are deleted. In this way, the algorithm creates nodes and clusters of nodes in the areas where the input signal distribution is highest. The algorithm stops when certain stopping criterion is met.

In our case each node cluster is a position of the robot. The robot therefore remembers the positions by placing nodes, or clusters of nodes in the R^n space of the GNG network. Each node or node cluster represents the information about the percept of the robot. The GNG algorithm does not have stopping criteria in our case, as the robot needs to learn the environment constantly.

3 Experiments and Evaluation Results

The system was implemented as in [5], on a PC and tested on a Lego Mindstorms NXT robot with rotating ultrasound sensor. The sensor readings were sent by the Lego NXT robot on the PC and the necessary calculations were done on the PC. This was done because of the limited memory and data structure organization that were available on the brick alone.

The system was tested in an indoor environment. The robot was given a task to explore an unknown environment by moving in a loop and memorizing the new and in the same time, recognizing the learnt positions. The robot also actively memorizes the sequence of actions it took in order to arrive at a certain position. After the learning process, the robot is expected execute a command sequence that would take it to a desired position.

The robot starts at an initial state. For each state the robot takes 12 readings from the ultrasound sensor in 12 different directions. These directions are then reordered according to the angle they were taken from. In absolute angles, from the initial direction (the direction of the initial pose) of the robot, the readings are taken from -180, -150, -120, -90, -60, -30, 0, 30, 60, 90, 120 and 150 angle degrees. If the robot direction is different from the initial direction, the relative angle of the reading directions is changed respectively. These values are then normalized for the GNG input in the presented approach and thus become the feature of the robots state (position). The GNG is then used to decide whether the robot's state is a new one or an old one, as in Fig 4. The obtained category identifies the state of the robot and is invariant of the robots direction angle. Which means that the sensor readings are ordered so that the array of the readings always start with the reading taken at -180 degrees and ends with 150 degrees in a global coordinate system for the environment. For simplicity the robot was allowed to rotate for 90 degrees in any direction and to move for a fixed length forward or backward.

Tests were made to evaluate the ability of the approach to recognize already visited states with the GNG algorithm. This allowed an estimation of the precision that the system has in learning the states. Due to the errors in the movement of the ultrasound sensor used, the robot tends to get different readings in the same state and map it as a different one. It also recognizes a new state as a state that it has already visited. For testing of the approach, the robot was ordered to move in a closed squared loop, as in Fig 5. The goal was for the robot to learn the new states and to recognize an old state when needed to. In the first loop the robot initializes the graph and remembers the states and the actions performed.

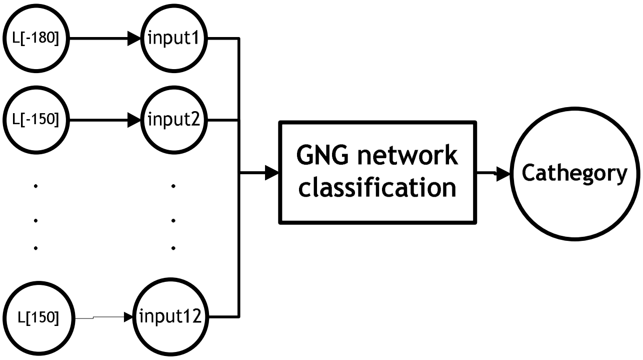


Fig. 4. Category determination using GNG network

The states are on a single step distance from each other. For each state the robot takes the measurements using the ultrasound sensor and the GNG network estimates the state that the robot is into. The robot was expected to recognize the states it has visited. The example labeling of the states is given in Fig 5. In this way we measured the precision of this method for the purpose of state recognition of the robot.

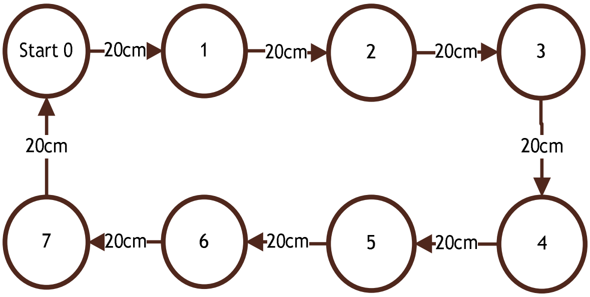


Fig. 5. States of the robot used for testing

The action planning for reaching a certain state using this method is a simple graph search algorithm that extracts the actions needed for reaching a certain position from a given position both represented as states.

The results of the tests are given in Fig 6. The experiment showed that both of the algorithms give similar results in the position estimation task. The FuzzyART is, however, slightly better than the GNG algorithm in the estimation of the state of the robot. The FuzzyART algorithm also tends to learn the positions faster, giving a stable position even after only one loop. The GNG algorithm learns the positions a bit slower and tends to recognize old states as new ones and continue to recognize them as the newer state in the next loops.

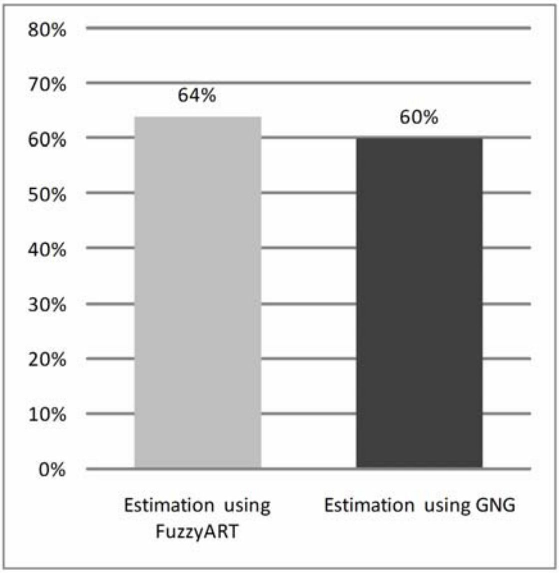


Fig. 6. Percentage of guessed categories

4 Conclusion

The obtained results showed errors in the state estimation. The reasons for these errors are the imprecision of the ultrasound sensor that is prone to miss some readings of the distance, especially if the object is a surface that has large angle with the ultrasound waves. This reason, combined with the angle errors of the rotation of the ultrasound sensor, tended to give imprecise readings that contained significant errors. These errors led the system to miss a state or to recognize two different states as the same category. The results shown in this paper are similar to the ones shown in [5]. The GNG approach however, was slightly less accurate.

These experiments showed that relying on a single ultrasound sensor for state estimation gives imprecise estimates. A good approach would be to use 12 or more different distance sensors attached in the appropriate angles in order to avoid the rotation errors of the single sensor. The nearly same error of both unsupervised learning algorithms confirmed the above statement.

Another reason for the obtained results is the ambiguity of the environment. The positions, which the robot was expected to recognize, were too close to each other and the errors of the sensor measurements of +-3cm were more than 10% of the distance between the positions, which also contributed to the error of the estimation. Furthermore the small distance between different positions narrows the gap between them in the state space also. Both the FuzzyART and the GNG networks occasionally tend to classify neighboring positions as the same one.

Another issue to be taken in consideration is that the robot learns the environment on the fly. This gives additional uncertainties in the state estimation. The main advantage of this approach however, is that the path planning of the mobile robot becomes a typical graph search problem that is easily solved.

In the future the same experiment would be repeated with 12 or more distance sensors including different types of equipment that would remove or at least minimize the errors that were made due to the imperfectness of the equipment used for these experiments.

Since the ultrasound sensors have shown to be imprecise and the learning of new positions and recognizing old ones, based on the data acquired from them, prone to errors, further consideration would be done to use different kinds of sensors for acquiring the environment data.

Acknowledgments

The work described in this article has been partially funded by the European Commission's Sixth Framework Programme under contract no. 029427 as part of the Specific Targeted Research Project XPERO ("Robotic Learning by Experimentation") and partially by the Ministry of Education and Science of Macedonia.

References

[1] Thrun, S., Burgard, W., Fox, D.: Probabalistic Robotics, 1st edn. The MIT Press, Cambridge (2005)

[2] Racz, J., Dubrawski, A.: Qualitative Pose Estimation Using An Artificial Neural Network. In: ICAR 1995 (1995)

[3] Mayer, G., Kaufmann, U., Kraetzschmar, G.: Neural robot detection in robocup. In: Wermter, S., Palm, G., Elshaw, M. (eds.) Biomimetic Neural Learning for Intelligent Robots. LNCS (LNAI), vol. 3575, pp. 349–361. Springer, Heidelberg (2005)

[4] Omidvar, O., Van der Smagt, P.: Neural Systems for Robotics. Academic Press, New York (1997)

[5] Lameski, P., Kulakov, A., Davcev, D.: Learning and position estimation of a mobile robot using FuzzyART neural network. In: IEEE/ASME International Conference on Advanced Intelligent Mechatronics (2009)

[6] Kulakov, A., Stojanov, G.: Structure, Inner Values, Hierarchies, and Stages: Essentials for Developmental Robot Architecture. In: Proceedings of the 2nd International Workshop on Epigenetic Robotics, pp. 63–70. Lund University Cognitive Studies (2002)

[7] Stojanov, G., Trajkovski, G., Kulakov, A.: Interactivism in artificial intelligence (AI) and intelligent robotics. New Ideas in Psychology 24, 163–185 (2006)

[8] Grossberg, S.: Adaptive Resonance Theory. Encyclopedia of Cognitive Science. Macmillan Reference Ltd., Basingstoke (2000)

[9] Carpenter, G.A., Grossberg, S., Rosen, D.B.: Fuzzy ART: Fast stable learning and categorization of analog patterns by an adaptive resonance system. Neural Networks 4, 759–771 (1991)

[10] Fritzke, B.: A self-organizing network that can follow non-stationary distributions. In: Gerstner, W., Hasler, M., Germond, A., Nicoud, J.-D. (eds.) ICANN 1997. LNCS, vol. 1327, pp. 613–618. Springer, Heidelberg (1997)

[11] Fritzke, B.: A growing neural gas network learns topologies. In: Advances in Neural Information Processing Systems (NIPS 1994), vol. 7, pp. 625–632. MIT Press, Cambridge (1995)

[12] Martinetz, T.M.: Competitive Hebbian learning rule forms perfectly topology preserving maps. In: ICANN 1993: International Conference on Artificial Neural Networks, Amsterdam, pp. 427–434. Springer, Heidelberg (1993)

Practical Method for Real-Time Path Planning and Optimization for Mobile Robots

Saso Koceski and Natasa Koceska

Faculty of Computer Sciences, University Goce Delcev, 2000 Stip, Macedonia
{koceski,njankova}@yahoo.com

Abstract. In the field of artificial intelligence and mobile robotics, calculating suitable paths, for point to point navigation, is computationally difficult. Maneuvering the vehicle safely around obstacles is essential, and the ability to generate safe paths in a real time environment is crucial for vehicle viability. A method for developing feasible paths through complicated environments using a baseline smooth path based on Hermite cubic splines is presented in this paper. A method able to iteratively optimize the path is also presented. This algorithm has been experimentally evaluated with satisfactory results.

Keywords: path planning, path optimization, Hermite cubic spline, obstacle avoidance, environment sensing.

1 Introduction

Path planning with motion modelling is an important and challenging task that has many applications in the fields of robotics, artificial intelligence (AI), virtual reality, autonomous agent simulation, etc. The basic task for the motion constraint path planning is to perform navigations from one place to another by coordination of planning, sensing and controlling whilst maintaining a smooth motion trajectory. For point to point navigation, calculating suitable paths is computationally difficult. Maneuvering the vehicle safely around obstacles is essential, and the ability to generate safe paths in a real time environment is crucial for vehicle viability.

Numerous motion planners consider the car-like vehicle as a three-dimensional system moving in the plane and subjected to constraints on the curvature in addition to the non-holonomic constraint of rolling without slipping. The pioneering work by Dubins [1] showed that the minimal length paths for a car-like vehicle consist of a finite sequence of two elementary geometrical components: arcs of circle and straight line segments. From then, almost all of the proposed motion planners compute collision-free paths constituted by such sequences [2]. As a result, the paths are piecewise C^2: they are C^2 along elementary components, but the curvature is discontinuous between two elementary components. To follow such paths, a real system has to stop at these discontinuity points in order to ensure the continuity of the linear and angular velocities. Continuous-curvature curve generation has become a key problem for on-going research in this area.

A few types of splines have been proposed to solve this problem. Gómez-Bravo et al [3] proposed a method for continuous curvature B-spline-based path planning for parking manoeuvres; Berlung et al [4] used the Bezier curve in path planning, having considered minimizing the square of the arc-length derivative of curvature along the curve. Shimizu et al [5] presented a method that uses clothoid curve for smooth path generation for mobile robot which is equipped with an omni-directional camera and a laser rangefinder. Scheuer [6] and Fraichard [7] also used clothoid curves in their vehicle control experiment. Unfortunately, clothoids do not have a closed form making the control of their shapes difficult and dangerous in the presence of obstacles. Other recent works [8], [9] adopted cubic splines in their trajectory generation algorithm. Later methods progressed to higher order polynomials [10]. However, previous work has mainly been focused on the static trajectory generation problem and on finding feasible solutions for 2D applications. All these solutions often require a great deal of computational power as they evaluate the entire path space [4], [11]. In a real time environment it is beneficial to directly compute feasible paths continuously to allow for variations in the environment, control error and unmodelled sensor error.

In this work a practical method for developing feasible paths, for nonholonomic car-like robot, through dynamic environments using a baseline smooth path based on Hermite cubic splines is presented. The developed method takes into consideration robot constraints. A method for adjusting and bending the spline to avoid obstacles is developed. This method is able to iteratively refine the path to more directly compute a feasible path and thus find an efficient, collision free path in real time through an unstructured environment. The efficiency of the proposed solution is evaluated in a custom, physics-based simulation environment provided by Open Dynamics Engine (ODE) [12]. In the simulation a simple car-like robot model equipped with 3D scanner and GPS has been developed. The generated motion path is smooth and has continuous curvature on the whole state space of the motion, thus satisfying the major requirements for the implementation of such strategies in real-time navigation.

2 Path Objectives

One of the objectives in this work is to study and develop a path planning algorithm for autonomous robot navigation or exploration in dynamic environments. The task can be divided into three parts: to plan a main path according to the pre-information, to keep tracking the difference between the map and the real environment, and then locally to amend the pre-designed path. This strategy can efficiently use the available information and reduce the re-planning time. It is supposed that the vehicle accepts a sequence of GPS waypoints used to define the high level mission. The robot's task then is to traverse through each waypoint.

The robot uses the 3D scanner to detect the surrounding environment and obtain the local information. The sensor is fixed to the robot body, and obstacles are mapped relative to the robot position and heading. The robot then uses the local information to generate the path to the destination.

Therefore, it is naive to pre-plan the entire robot path from the outset. Thus, the path is re-planed from the current position and heading, and using the most current obstacle map. The path is also limited to only look ahead past the next waypoint. In this way, the next turn will be feasible and once executed, a new path will be generated to maneuver the vehicle into a suitable position for the following turn. For this reason, only three consecutive waypoints are used at a time, the most recently passed, and the following two points. This approach can be also used in the cases of unknown environments, where no pre-information is available before the path planning algorithm has been executed.

3 Path Planning Using Hermite Cubic Splines

In this paper the implementation of Hermite cubic splines as a tool for path planning is adopted. The mathematics involved in creating splines (which are piecewise polynomial functions), allow easy construction of smooth paths through a given, finite set of control points. Given *N+1* control points, and knowing the robot starting and goal position and orientation, a series of *N* spline segments are generated, with the three-order polynomial functions of variable *t* ($t \in [0, 1]$), to traverse these points, as:

$$\begin{bmatrix} X(t) \\ Y(t) \end{bmatrix} = \begin{bmatrix} a_x & b_x & c_x & d_x \\ a_y & b_y & c_y & d_y \end{bmatrix} \begin{bmatrix} t^3 \\ t^2 \\ t \\ 1 \end{bmatrix} . \tag{1}$$

where *X(t), Y(t)* are the coordinates of any point on the cubic spline, a_x, b_x, c_x, d_x and a_y, b_y, c_y, d_y are the coefficients to determine.

As the first derivative of the path is proportional to the vehicle heading, a non-continuous derivative would result in an infeasible path for this type of vehicle, but the second derivative is proportional to the vehicle steering angle and any discontinuities would force the vehicle to stop at each control point to adjust its steering. By creating a path with continuous derivatives, a smooth vehicle control, to remain in motion throughout the vehicle path, is guaranteed.

For these *N* segments of cubic spline, the required number of equations is (*8N*) in order to solve out all the coefficients. The known conditions are:

- the initial and final position and the robot orientation in these points;
- the continuity of positions at (*N-1*) control points;
- the continuity of 1st derivatives at (*N-1*) control points;
- the continuity of 2nd derivatives at (*N-1*) control points;

The total number of know conditions is (*6N+2*) which remove (*6N+2*) degrees of freedom from the 8N ones. The number of remained degrees of freedom is 2(*N-1*). This number is exactly the same as the unknown *x*-, *y*-coordinates of (*N-1*) control points. One set of points determines one path. By searching for the suitable control points, a feasible cubic splines path can be determined.

Considering just the local parameterization of the i^{th} cubic spline sequence in only the *x* direction, we will have:

$$x_i(t) = a_i + b_i t + c_i t^2 + d_i t^3 \ . \tag{2}$$

Any cubic equation can be used to construct a cubic spline by identifying the constants a_i, b_i, c_i and d_i; however, the natural Hermite cubic polynomial has a unique property where it satisfies all four of the following boundary conditions:

$$\begin{aligned} & x_i(0) = a_i, \\ & x_i(1) = a_i + b_i + c_i + d_i, \\ & x'_i(0) = D_i = b_i, \\ & x'_i(1) = D_{i+1} = b_i + 2c_i + 3d_i \end{aligned} . \tag{3}$$

By re-arranging and writing $x_i(1) = x_{i+1}$, we will have:

$$\begin{aligned} & a_i = x_i \\ & b_i = D_i \\ & c_i = 3(x_{i+1} - x_i) - 2D_i - D_{i+1} \\ & d_i = -2(x_{i+1} - x_i) + D_i + D_{i+1} \end{aligned} . \tag{4}$$

If equate the 2nd derivatives p=1 for the (i-1)th segment, p=0 for the ith segment we will have:

$$\begin{aligned} & 2c_{i-1} + 6d_{i-1} = 2c_i \\ & \Rightarrow 2[3(x_i - x_{i-1}) - 2D_{i-1} - D_i] \\ & + 6[-2(x_i - x_{i-1}) + D_{i-1} + D_i] = \\ & 2[3(x_{i+1} - x_i) - 2D_i - D_{i+1}] \\ & \Rightarrow D_{i-1} + 4D_i + D_{i+1} = 3(x_{i+1} - x_{i-1}) \end{aligned} . \tag{5}$$

At the first point, $x''_0(0) = 0$. So $c_0=0$ and $3(y1\text{-}y0)\text{-}2D_0\text{-}D_1=0$, so:

$$2D_0 + D_1 = 3(x_1 - x_0) \ . \tag{6}$$

Similarly for the end section $x''_N(1) = 0$. So $2c_m+6d_m=0$ from which

$$D_{N-1} + 2D_N = 3(x_N - x_{N-1}) \ . \tag{7}$$

Gather all this together the solution of the spline path in the matrix form, will be:

$$\begin{bmatrix} 2 & 1 & & & & & \\ 1 & 4 & 1 & & & & \\ & 1 & 4 & 1 & & & \\ & & & \vdots & & & \\ & & & & 1 & 4 & 1 \\ & & & & & 1 & 2 \end{bmatrix} \begin{bmatrix} D_0 \\ D_1 \\ D_2 \\ \vdots \\ D_{N-1} \\ D_N \end{bmatrix} = \begin{bmatrix} 3(x_1 - x_0) \\ 3(x_2 - x_0) \\ 3(x_3 - x_1) \\ \vdots \\ 3(x_N - x_{N-2}) \\ 3(x_N - x_{N-1}) \end{bmatrix} . \tag{8}$$

As mentioned previously, obstacles are mapped from various sensors and stored into the map Fig. 1. It should be noted that each obstacle must be inflated by at least half the width of the vehicle to guarantee that a collision does not occur (in the future discussion, talking about obstacles, the inflated obstacles will be considered). Considering this, and knowing the start and goal positions and orientations, for each pair of segments, the initial path composed of Hermite cubic splines (dashed line in Fig. 1) is constructed.

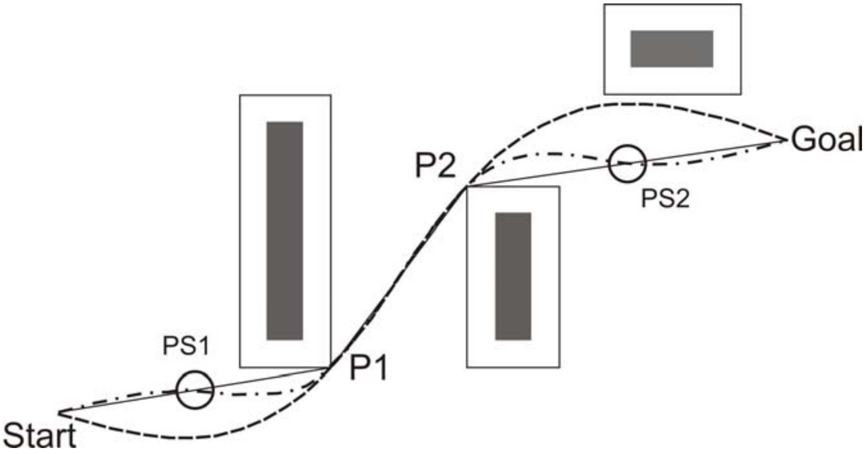


Fig. 1. Spline path passing through the initial control points (dashed). The introduction of additional control points can keep the initial path away from obstacles and decrease the path length (dot-dash).

4 Path Refinement and Optimization

After the construction of the initial spline path it is useful to include additional points along the straight line path, connecting the initial control points.

The quantity and number of these points depend greatly on the relationship between the individual path lengths and the obstacles distribution. By including additional points, the initial path is kept closer to the nominal straight line path and it is therefore shorter. These points are only used for the initial path optimization and will not be rigid constraints in the final path. For demonstration, in the previous example depicted in Fig. 1, we have included two additional points PS1 and P2G which are midpoints of the first and third path segments respectively. The path containing these points is depicted as a dot dash line. As one may observe, addition of such points can iron the path and can be very useful in the environments where the robot has to traverse a corridor or a narrow passage. In order to evaluate the quality of the paths, the following optimization function is introduced:

$$f = \frac{l}{l_m} + \left(\frac{\alpha}{d_{\min}} \right)^2 + k \cdot \frac{Rr_{\min}}{R_{\min}} \ . \tag{9}$$

where the l_{min} is the Euclidian distance between Start and Goal, α is a weight constant of the distance from robot to obstacles, k is a weight constant of the minimum radius for robot driving, l is the total path length, d_{min} is the minimal distance between any point on the path and the obstacles, given by:

$$d_{\min} = \min_{o \in O} \min_{P \in Path} \sqrt{(X_P - X_o)^2 + (Y_P - Y_o)^2} \ . \tag{10}$$

where P denotes any point on the path, O is the set of all obstacles in the environment. R_{min} is the minimum radius along the whole path and Rr_{min} is the minimum turning radius, the robot can deal with. By minimizing this function, it is possible to shorten the path length, keep the robot as far as possible away from obstacles and smooth enough. Eventually the optimal path is a compromise of all the requirements. This optimization function strictly penalizes the trajectories that cause collision with the obstacles.

If it has been determined that the path collides with an obstacle, the spline has to be manipulated to avoid that obstacle. The proposed method is based on adding of an additional control point to the spline segments between the intersection points, to guide the path around the obstacle. As multiple collisions may have occurred, we will have a list of collision points. In order to minimize the computational time this method first calculates the convex hull of the obstacle. Then it calculates the pair of intersection points of the spline path and the convex hull of the obstacle. First point corresponds to the entry point of the spline into the convex hull and the second one, to the exit point, where the spline goes out of the hull area.

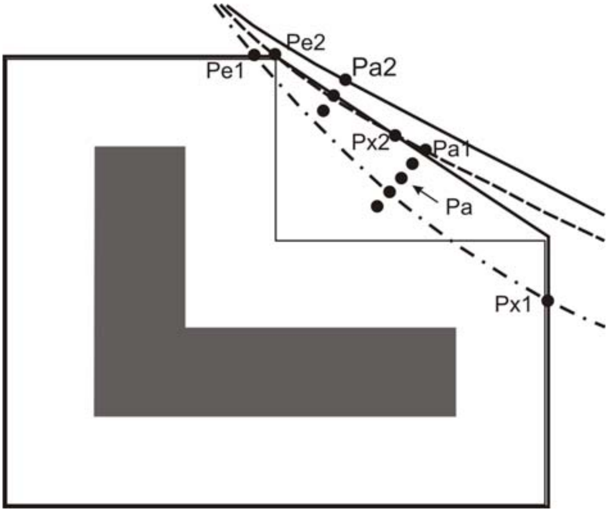


Fig. 2. Initial spline path (dot dashed line) intersects the computed convex hull (bolded line polygon) around the obstacle (filled grey). Modified path after the first optimization is presented as dashed line, the final save path obtained after the second optimization is depicted as solid line.

The method then adds an additional point, P_a, on the segment, between the entry point (P_e) and the exit point (P_x)(Fig. 2). Then this point is moved perpendicularly to the segment $\overline{\mathrm{P_e\,P_x}}$, by a predefined small distance (d_m) at each iteration, in both directions. The point P_a is continually updated, evaluated and checked against the obstacle's hull, until P_a is free of collision. This means that a path through P_a will no longer collide with the obstacle at that point. P_a is than added to the list of path control points. A new spline (depicted as dashed line in Fig. 2) is computed through these points still having the desired characteristics but also passing through the new point. For each new spline the process is repeated. If the new spline still intersects the hull (P_{e2} ,P_{x2}), the collision will be shorter and closer to the edge of the hull. Thus the algorithm will continue to displace the spline around the remaining portion of the hull. Because the points are fit with cubic functions, a large number of control points within close proximity to each other can cause large deviations in the path and increasing of its curvature. To avoid this, when adding a new point, any other control point within a given radius is removed. Therefore, in the example in Fig. 2, the new added point P_{a2}, obtained in the second optimization step, replaces the point P_{a1}, calculated in the previous step.

5 Simulation Experiments and Results

In order to test the path planning algorithm a physics-based simulation environment provided by ODE has been developed. For the scope of this work a four wheels mobile robot model which has a similar structure to the normal car is considered (i.e. two front steering wheels and two driven rear wheels). The developed moving robot model has two degrees of freedom and the dynamics of the model can be represented as a set of motion's equations in terms of mass, accelerations and steering angles as well as external force conditions, such as ground friction. The dynamics of a moving robot must follow the basic law of motion dynamics, which may be represented as a set of general ordinary differential equations in the form:

$$\begin{aligned} \frac{d^2X}{dt^2} &= \hat{f}\left(X,\dot{X},\delta\right) + \Delta(\delta) \\ \Delta(\delta) &= f\left(X,\dot{X},\delta\right) - \hat{f}\left(X,\hat{\dot{X}},\delta\left(\hat{a},\hat{\theta}\right)\right) . \end{aligned} \tag{11}$$

where, $\hat{\dot{X}},\hat{a},\hat{\theta}$ are approximate values of motion velocity, acceleration and direction of motion (i.e. a steering angle) respectively, and the motion control δ is a function of acceleration a and moving direction θ. A desired or predicted motion state of the moving object is pre-estimated by a set of approximate functions according to the state of moving object and the environment conditions related to the surrounding obstacle–space. The actual motion track is then computed. The difference between the predicted motion and the actual motion will be used for estimating the control input to mobile robot system.

All robot wheels have the same diameter and two rear wheels are conventional fixed wheels on the same axle and two front wheels are centered orientation wheels. The wheels are modeled using ODE's basic collision primitive, cylinder, and they are connected to the base body using motorized hinge joints with a horizontal rotational axis (vertical rotation plane). ODE also provides the possibility to set a desired velocity with a maximum force for each wheel.

The steering angle can be expressed as:

$$\phi = a\tan(L/R) \ . \tag{12}$$

where L is the robot length and R is the distance from the middle point P_M of the rear wheels to the instantaneous center of curvature (ICC). The axes of all four wheels pass through the ICC during the driving. Due to the fixed rear wheels, the robot is not permitted to change its orientation on the spot like the omni-directional robots.

The nonholonomic constraint for this kind of robot is expressed as:

$$\tan\theta = \dot{y}_{P_M} / \dot{x}_{P_M} \ . \tag{13}$$

The angle θ stands for the orientation of the robot frame. This constraint says that the direction of the translational velocity is the tangent direction of the path. The physical meaning of this constraint is that there is no possible motion in the axial direction. The robot is also equipped with 3D scanner and GPS sensor simulation models. Simple testing environment with rectangular shape and dimensions 20x20m, cluttered with random positioned obstacles with random shape and dimensions, was created, start and goal robot positions and orientations were set up (Fig. 3).

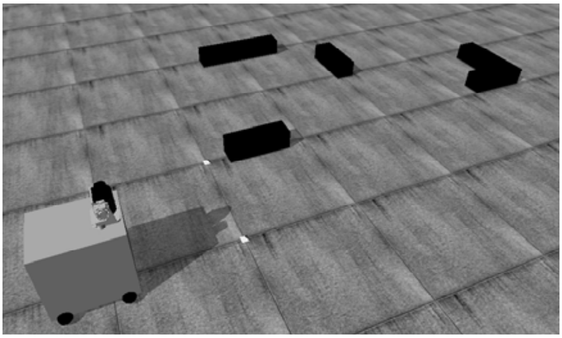


Fig. 3. 3D Simulation environment

One simple map with the obstacles and safety margins around, as well as the spline path computed in real-time using the algorithm presented in this work are depicted in Fig. 4a. The optimal path in this case is calculated in real time after two iterations. The efficiency of many path planning algorithms decreases dramatically in spaces with narrow passages. The proposed algorithm is also tested in the case of a simple narrow passage between two obstacles and the collision free path is found after a single iteration (Fig. 4b).

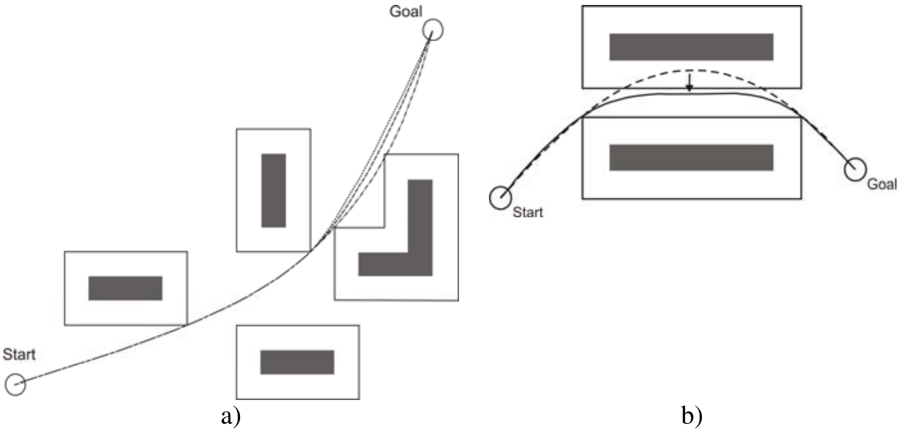


Fig. 4. a) Calculated spline path after two optimization steps, b) Final path through a narrow passage between two obstacles obtained after a single optimization step.

The path, as defined by the algorithm, ensures safe passage of the vehicle through its environment and contains all necessary information. The steering angle can be derived from the second derivative of the path. Feedback is applied to keep the vehicle on the spline path. The average end-position error in the process of path following, along the whole path, in all cases is less than 30 mm.

6 Conclusion and Future Work

A novel motion constraint path planning approach for real-time navigation of mobile robots is proposed in this paper. The algorithm is able to produce a collision-free, time-optimal smooth motion trajectory. A 3D simulation has been conducted and the result is quite promising. The simulation result is quite satisfactory. The next step of our research is to refine the algorithm and make the method more robust in complex environments.

References

1. Dubins, L.E.: On curves of minimal length with a constraint on average curvature and with prescribed initial and terminal positions and tangents. Amer. J. Math. 79, 497–516 (1957)
2. Lamiraux, F., Laumond, J.-P.: Smooth Motion Planning for Car-Like Vehicles. IEEE Trans. of Robotics and Automation 17(4), 498–502 (2001)
3. Gómez-Bravo, F., Cuesta, F., Ollero, A., Viguria, A.: Continuous curvature path generation based on ß-spline curves for parking manoeuvres. Robot. Auton. Syst. 56(4), 360–372 (2008)
4. Berglund, T., Jonsson, H., Soderkvist, I.: An obstacle-avoiding minimum variation b-spline problem. In: Proc. of the 2003 International Conference on Geometric Modeling and Graphics (2003)

5. Shimizu, M., Kobayashi, K., Watanabe, K.: Clothoidal Curve-based Path Generation for an Autonomous Mobile Robot. In: Proc. of the 2006 International Joint Conference SICE-ICASE, pp. 478–481 (2006)
6. Scheuer, A., Fraichard, T.: Planning Continuous-Curvature Paths for car-Like Vehicles. In: IEEE-RSJ Int. Conf. on Intelligent Robots and Systems, vol. 3, pp. 1304–1311 (1996)
7. Fraichard, T., Ahuactzin, J.M.: Smooth Path Planning for Cars. In: IEEE Int. Conf. on Robotics and Automation (2001)
8. Nagy, B., Kelly, A.: Trajectory Generation for Car-Like Robots Using Cubic Curvature Polynomials. In: Field and Service Robots 2001, Helsinki, Finland (2001)
9. Saska, M., Macas, M., Preucil, L., Lhotska, L.: Robot Path Planning using Particle Swarm Optimization of Ferguson Splines. In: ETFA 2006 Proceedings [CD-ROM], Piscataway. IEEE, Los Alamitos (2006)
10. Thompson, S., Kagami, S.: Continous curvature trajectory generation with obstacle avoidance for car-like robots. In: Proceedings of the 2005 International Conference on Computaional Intelligence for Modelling, Control and Automation and International Intelligent Agents, Web Technologies and Internet Commerce (2005)
11. shiller, Z., Gwo, Y.-R.: Dynamic motion planning of autonomous vehicles. IEEE Transactions on Robotics and Automation 7(2), 241 (1991)
12. Smith, R.: Open Dynamics Engine - ODE (2008), `http://www.ode.org`

Protein Classification Based on 3D Structures and Fractal Features

Georgina Mirceva[1], Zoran Dimov[2], Slobodan Kalajdziski[1], and Danco Davcev[1]

[1] Faculty of Electrical Engineering and Information Technologies, Skopje, Macedonia
{georgina,skalaj,etfdav}@feit.ukim.edu.mk
[2] Microsoft Corporation, Vancouver BC, Canada
zodimov@microsoft.com

Abstract. To understand the structure-to-function relationship, life sciences researchers and biologists need to retrieve similar structures and classify them into the same protein fold. In this paper, we propose a 3D structure-based approach for efficient classification of protein molecules. Classification is performed in three phases. In the first phase, we apply fractal descriptor matching as a filter. Then, protein structures which satisfy the fractal and radius tolerance are classified in the second phase. In this phase, 3D Fourier Transform is applied in order to produce rotation invariant descriptors. Additionally, some properties of primary and secondary structure are taken. In the third phase we use k nearest neighbor classifier. Our approach achieves 86% classification accuracy with applying fractal filter, and 92% without fractal filter. It is shown that fractal filter significantly shorten the classification time. Our system is faster (seconds) than DALI system (minutes, hours, days), and we still get satisfactory results.

Keywords: Protein classification, fractal descriptor, 3D Discrete Fourier Transform, DALI.

1 Introduction

The structure of a protein molecule is the main factor which determines its chemical properties as well as its function. All information required for a protein to be folded in its natural structure is coded in its amino acid sequence. Therefore, the 3D representation of a residue sequence and the way this sequence folds in the 3D space are very important. The 3D protein structures are stored in the world-wide repository Protein Data Bank (PDB) [1], [2] which is the primary repository for experimentally determined proteins structures. With the technology innovation the number of 3D protein structures increases every day. As the number and variety of proteins continue to grow there has been an increasing interest in applications to help navigate through these large databases. There are various methods for protein retrieval and classification according to their structure.

In [3], the ray-based method as a 3D-model retrieval technique is introduced. Silhouette, depth buffer, volume and voxel-based feature vector are presented in [4].

The geometric hashing method to perform protein surface matching to identify similar binding sites is presented in [5]. Two techniques, α-hull and 3D reference frames, are adopted to reduce the complex computation.

In [6], another protein structure retrieval system is proposed. They constructed an indexing structure to avoid exhaustively chain structure alignments. Relevant signatures have been extracted from 2D distance matrices.

Protein classification plays a central role in understanding the function of a protein molecule. With the rapid increase in the number of new proteins, the need for automated and accurate methods for protein classification is increasingly important.

In [7], a new scheme for automatic classification of 3D protein structures is presented. It is a dedicated and unified multiclass classification scheme. A nearest neighbor classifier has been adopted. A filter-and-refine scheme is used.

In [8], nine different protein classification methods are included for the performance analysis. The nine methods used are the profile-HMM, support vector machines (SVMs) with four different kernel functions, SVM-pair wise, SVM-Fisher, decision trees and boosted decision trees.

In this paper, we present a system for classifying protein tertiary structures. First, we apply fractal descriptor [9] matching as a filter. Then, protein 3D structures with satisfactory fractal and radius tolerance are classified in the second phase. We have adopted the method given in [4] to extract the geometry descriptor. Additionally, some features of primary and secondary structure of protein molecule are taken as in [10], thus forming better integrated descriptor.

There are many algorithms used for protein classification as Naive Bayesian classifier, nearest neighbor classifier, decision trees and so on. In our approach, we have used *k* nearest neighbor classifier [11]. The evaluation of the classification algorithm is made according to the DALI method [12]. We provide some experimental results concerning the usage of the fractal filter.

The proposed research approach is given in section 2; while in section 3 the experimental results and evaluation of the system are presented. Section 4 concludes the paper and gives some future work directions.

2 Our Approach

The information about protein structure is stored in PDB files. The PDB files are stored in the Protein Data Bank (PDB) [2], which is the primary depository of experimentally determined protein structures. PDB files contain information about primary, secondary and tertiary structure of protein molecules. We will make the observation that one protein is totally described by the arrangement of its atoms in the 3D Euclidian space.

Our goal is to provide a system which provides structural classification of protein structures. Phases of our classification system are illustrated on Fig. 1. The user uploads the PDB file of the comparing protein. The information from PDB file is processed and fractal descriptor [9] is extracted. Fractal descriptor is compared with fractal descriptors of the proteins in the database according to Euclidean distance. After some processing (triangulation, voxelization), 3D descriptor for the query protein is extracted. The 3D descriptor is compared with 3D descriptors of proteins

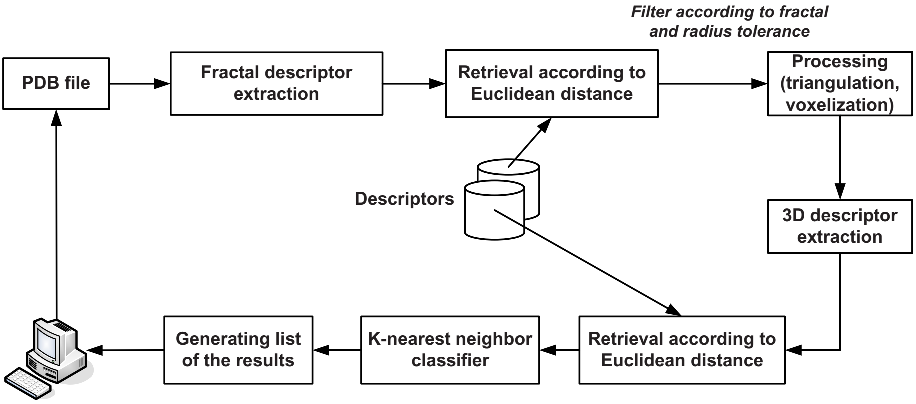

Filter according to fractal and radius tolerance
PDB file
Fractal descriptor extraction
Retrieval according to Euclidean distance
Processing (triangulation, voxelization)
Descriptors
3D descriptor extraction
Generating list of the results
K-nearest neighbor classifier
Retrieval according to Euclidean distance

Fig. 1. Phases of our classification system

which satisfy the fractal and radius tolerance (defined by the user). Then, protein is classified according to *k*-nearest neighbor method [11] and list of results is shown.

2.1 Volume Fractal Descriptor

We have used the fractal descriptor proposed in [9] as a filter. The 3D space is divided in elementary cubes with side length of l (the scaling factor). Then we compute the volume of the proteins 3D shape $V(l)$ with the box counting method, by calculating the ratio of the number of non empty elementary cubes. For effective and robust results, l should be chosen small enough, because the volume $V(l)$ would be invariant of l. We can obtain the volume fractal dimension of protein as in (1).

$$D = \lim_{l \to 0} \frac{log(V(l))}{log(l)} \tag{1}$$

For the software implementation for computing (1), we can use the method for linear regression taking the line equation $y=ax+b$ as model. Let the first N discrete samples from this line are (x_i,y_i), $1 \le i \le N$, where $x_i=log(l_i)$, $y_i=log(V(l_i))$. Plots of the points (x_i,y_i), $1 \le i \le N$, for protein 7XIM are depicted on Fig. 2. The fractal dimension would be the slope a of the line $y=ax+b$.

The experimental results showed that the fractal dimension of protein, as unique descriptor feature, cannot be robust describer of its geometric features. Two proteins can have very close fractal dimension, but one of them can be spatially much more nearer. Thus, we use the proteins radius as second feature. With the radius R of protein we represent the radius of the smallest sphere that encloses whole protein 3D shape. With only these two features an effective protein descriptor is build. This descriptor is no memory consuming. We use the L_2 norm, so matching is very fast.

A protein fractal descriptor on a smaller dataset plotted as points (D, R) is shown on Fig. 3. The red point is descriptor of query protein and it has four descriptors

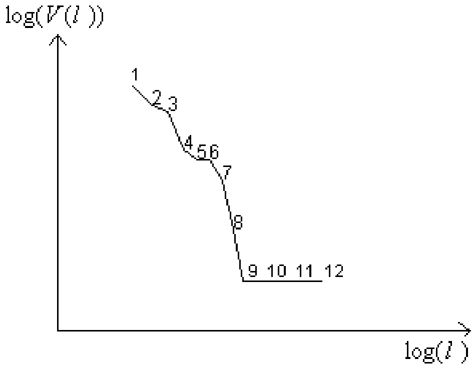


Fig. 2. Plots for the protein 7XIM

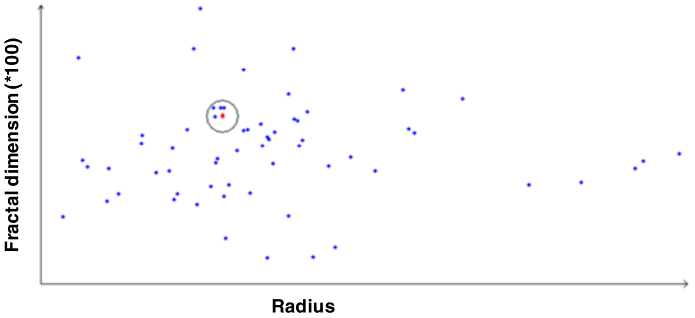


Fig. 3. Plotting descriptors as points

which satisfy the predefined threshold (fractal and radius tolerance). Fractal and radius tolerance are defined by the user.

The fractal descriptor is not robust enough to retrieve the proteins that best match to the query protein, but using the *k*-nearest neighbors we can find the set of proteins which potentially match the query. In other words, we can reduce the database for up to 90% and then use some more robust methods, which are more time and memory consuming.

2.2 Voxel Based Descriptor

We have used the voxel-based algorithm presented in [4] to extract the geometry descriptor and the Euclidean distance as a metric for comparison. In [4], this algorithm is proposed for any kind of objects (airplanes, cars etc.). In this paper we have used this algorithm for building protein structure classification system.

Since the exact 3D position of each atom and its radius are known (according to PDB file), it may be represented by a sphere. First, we perform triangulation in order to build a mesh model which presents the protein structure. The surface of each sphere is triangulated. In this way, a sphere consists of a small set of vertices and a set of connections between the vertices. Finally, a protein is comprised of a set of spheres, along with the corresponding vertices and the connections among them. Then, the centre of mass is calculated and the protein is translated so the new centre of mass is at the origin. The distance d_{max} between the new origin and the most distant vertex is computed and protein is scaled, so d_{max} =1. In this way, we provide translation and scale invariance.

After triangulation, we perform voxelization. Voxelization transforms the continuous 3D-space, into discrete 3D voxel space. The voxelization proceeds in three steps: discretization, sampling, and storing. Discretization divides the continuous 3D-space into voxels. With sampling, depending on positions of the polygons of a 3D-mesh model, to each voxel v_{abc} a value is attributed equal to the fraction of the total surface area S of the mesh which is inside the region μ_{abc} (2).

$$v_{abc} = \frac{area\{\, \mu_{abc} \cap I \,\}}{S}, \; 0 \leq a,b,c \; \leq N - 1. \tag{2}$$

Each triangle T_j of a model is subdivided into p_j^2 coincident triangles each of which has the surface area equal to $\delta = S_j / p_j^2$, where S_j is the area of T_j . If all vertices of the triangle T_j lie in the same cuboid region μ_{abc}, then we set p_j = 1, otherwise we use (3) to determine the value of p_j, as in [4].

$$p_j = \left\lceil \sqrt{p_{min} \frac{S_j}{S}} \right\rceil \tag{3}$$

For each newly obtained triangle, the center of gravity G is computed, and the voxel μ_{abc} is determined. Finally, the attribute v_{abc} is incremented by δ. The quality of approximation is set by the parameter p_{min}. In our implementation we have set p_{min} =32000.

The information contained in a voxel grid can be processed further to obtain both correlated information and more compact representation of voxel attributes as a feature. We applied the 3D Discrete Fourier Transform (3D-DFT) to obtain a spectral domain feature vector which also provides rotation invariance of the descriptor. A 3D-array of complex numbers $F = [f_{abc}]$ is transformed into another 3D-array by (4).

$$f'_{pqs} = \frac{1}{\sqrt{MNP}} \sum_{a=0}^{M-1} \sum_{b=0}^{N-1} \sum_{c=0}^{P-1} f_{abc} e^{-2\pi j(\,ap/M + bq/N + cs/P\,)} \tag{4}$$

Since we apply the 3D-DFT to a voxel grid with real-valued attributes, we shift the indices so that (a; b; c) is translated into (a–M/2; b–N/2; c–P/2). Let M=N=P and we introduce the abbreviation (5).

$$\upsilon'_{a-M/2,b-N/2,c-P/2} \equiv \upsilon_{abc} \tag{5}$$

Thus, the origin (0; 0; 0) is shifted to (N/2; N/2; N/2). We take magnitudes of low-frequency coefficients as components of the vector. Since the 3D-DFT input is a real-valued array, the symmetry is present among obtained coefficients, so the feature vector is formed from all non-symmetrical coefficients (6).

$$1 \leq |p| + |q| + |s| \leq k \leq N/2 \tag{6}$$

We form the feature vector by the scaled values of f'_{pqs} by dividing by $|f'_{000}|$. This vector presents geometrical properties of the protein.

Additionally, characteristic attributes of the primary and secondary structure of the protein molecules are extracted, forming attribute-based descriptor vectors as in [10]. More specifically, concerning the primary structure, the ratios of the amino acids' occurrences and hydrophobic amino acids ratio are calculated. Concerning the secondary structure, the ratios of the helix types' occurrences, the number of Helices, Sheets and Turns in a protein are also calculated. These features and the weights assigned to them are listed in Table 1.

Table 1. Structural features and their weights

Secondary structure features	Weight (%)
Ratios of Helix types	1
Number of HELICES	1
Number of SHEETS	1
Number of TURNS	1
Primary structure features	**Weight (%)**
Hydrophobic residue ratio	6
Residue ratios	90

Hydrophobic amino acids tend to be in the centre of the protein 3D structure, while hydrophilic amino acids tend to be on the surface of the protein 3D structure. Thus hydrophobic amino acids ratio very well gives primary information about 3D structure of the protein. That's why a significant weight is assigned to this feature.

In the process of retrieval, voxel descriptors are compared according to their Euclidean distance.

The geometrical descriptors are compared in pairs by using (7), as in [4].

$$D_G = \min_{\alpha \in R} d_p(f_g^{'}, \alpha f_g^{''}) = \min_{\alpha \in R} \left\| f_g^{'} - \alpha f_g^{''} \right\|_p \tag{7}$$

For $p = 2$, the parameter α is computed by using (8).

$$\alpha = f_g^{'} * f_g^{''} / f_g^{''2} \tag{8}$$

The structural similarity is evaluated by (9), where additionally different weights (see Table 1) to the attributes were assigned.

$$D_S = \sqrt{\sum_{i=1}^{34} W_i [f_s^{'}(i) - f_s^{''}(i)]^2} \tag{9}$$

The overall similarity is determined by (10). As it can be seen from (10), our algorithm is mainly based on geometrical features (90%) rather than structural features (10%).

$$D= k_1D_G+ k_2D_S \; , \quad k_1=90\%, \quad k_2=10\% \tag{10}$$

By using the overall similarity measure, the distance between descriptor of comparing protein and descriptors of proteins which satisfy the fractal and radius tolerance are calculated, and a list of results according to Euclidean distance is returned.

2.3 Distance-Weighted k Nearest Neighbors Classifier

There are many algorithms used for protein classification as Naive Bayesian classifier, nearest neighbor classifier, decision trees and so on. In this approach we have used *k* nearest neighbor classifier [11] in the classification of proteins. As the name indicates, *k*-nearest neighbors of the query protein *q*, are used to determine the class of *q*. Different weights are assigned to the *k* neighbors based on their distance from the query protein, namely inverse square of the distances is used as weight. The effectiveness depends on the number *k* as well as on the weighting of the *k* neighbors.

We have used (11) in order to classify the query protein, where function *f*(*i*,*Cl*) is a binary function which present if the *i*-th most similar protein belongs to class *Cl*.

$$S_{Cl} = avg[\frac{1}{d_i^{\,2}} * f(i,Cl)] \; , \quad i=1,\ldots, k \tag{11}$$

3 Experimental Results

We have implemented a web-based system for protein classification. The dataset consists of 1065 proteins from 26 classes from the FSSP/DALI database [13]. Some experimental results for classifying the protein transferase are shown on Fig. 4. The first protein that is shown (n. 1) is the most similar protein of the winning class; the second one is the most similar protein from the second winning class and so on. Only classes to whom the first k proteins belong are shown.

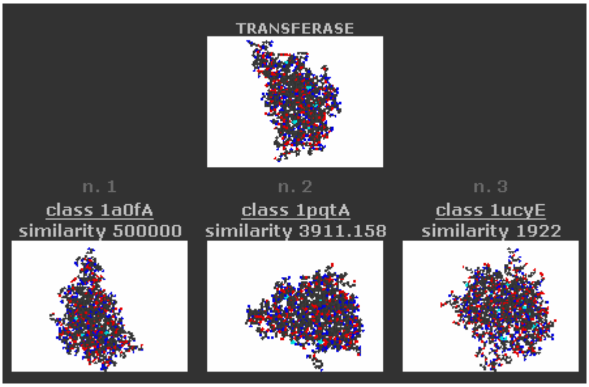


Fig. 4. Experimental results for classifying the protein transferase

We have compared our results on a test set of 50 randomly selected proteins with the results of DALI system and we got more than 92% classification accuracy without applying fractal filter and 86% classification accuracy with applying fractal filter. From the performance point of view, the analysis showed that our system is faster (takes few sec) than the DALI system which lasts much longer (minutes, hours), and we still get satisfactory results.

First, we examined the classification time taken from our system with using fractal filter against classification time taken by the DALI system. The ratios between the times taken from our system by using fractal filter and the DALI system are presented on Fig. 5. As it can be seen, DALI takes much longer than our system. For example, for the first selected protein, DALI system response time was 91 sec., while the response time of our system was 1.87 sec., so the ratio 1.87/ (1.87+91) = 2% corresponds to the first experiment on Fig. 5. Corresponding ratios for the other proteins were calculated in the same way.

In addition, we investigated the influence of using fractal filter. For the same test protein, by using fractal filter, our system's response time was 1.875 sec, while in the case without fractal filter the response time was 6.26 sec. The ratio of classification time with and without applying fractal filter is shown on Fig. 6. It can be seen that classification lasts less time when we use the fractal filter. The difference between classification times of our system with and without fractal filter will significantly increase as the number of proteins in the database increases.

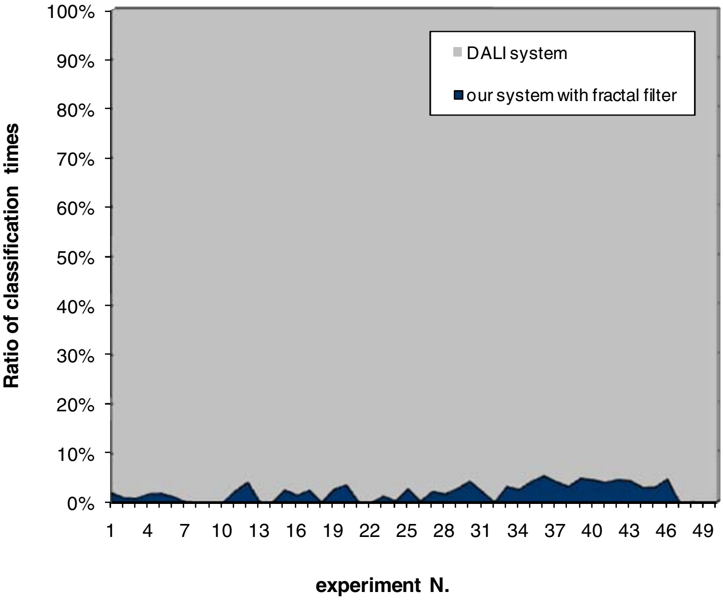


Fig. 5. The ratio of classification times of our system by using fractal filter and DALI system

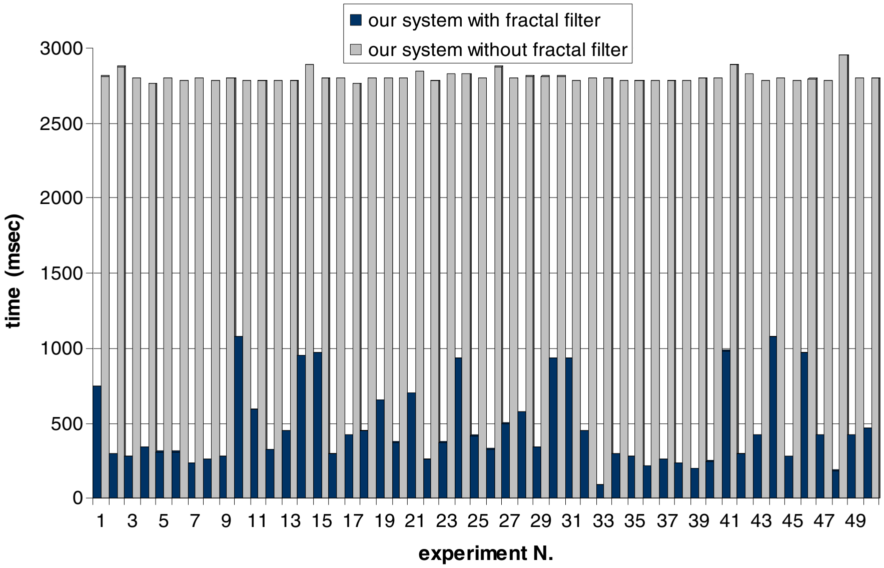


Fig. 6. The ratio of classification times of our system with and without fractal filter

4 Conclusion

We have presented a system for protein molecules classification by using information about their primary, secondary and tertiary structures. We also applied fractal method as a filter, in order to obtain faster classifier.

A part of the FSSP/DALI database, which provides a structural classification of the proteins, was used to evaluate the classification. The results show that our system achieves more than 92% classification accuracy without applying fractal filter and 86% classification accuracy with applying fractal filter, while it is much simpler and faster (few sec) than the DALI system (minutes, hours).

We also provide some experimental results concerning the usage of the fractal filter. It was shown that fractal filter significantly shorten the classification time, that will be much obvious on larger dataset.

Our future work will be concentrated on increasing the efficiency of the system by investigating new descriptors and incorporating additional characteristics in the descriptors. Also, more sophisticated classifiers can be used that would lead to faster and more accurate system.

References

1. Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N., Bourne, P.E.: The Protein Data Bank. Nucleic Acids Research 28, 235–242 (2000)
2. Protein Databa Bank, `http://www.rcsb.org`

3. Vranic, D.V.: An improvement of Ray-Based Shape Descriptor. In: Wittig, W., Paul, S. (eds.) 8th Leipziger Informatik-Tage (LIT'2M), Leipzig, Germany, pp. 55–58. HTWK Leipzig (2000)
4. Vranic, D.V.: 3D Model Retrieval. Ph.D. Thesis. University of Leipzig (2004)
5. Chen, S.C., Chen, T.: Protein Retrieval by Matching 3D Surfaces. In: Genomic Signal Processing and Statistics (GENSIPS 2002), Raleigh, North Carolina, USA (2002)
6. Chi, P.H., Scott, G., Shyu, C.R.: A Fast Protein Structure Retrieval System Using Image-Based Distance Matrices and Multidimensional Index. In: Fourth IEEE Symposium on Bioinformatics and Bioengineering (BIBE 2004), Taichung, Taiwan, pp. 522–532 (2004)
7. Aung, Z., Tan, K.L.: Automatic 3D Protein Structure Classification without Structural Alignment. Journal of Computational Biology 12(9), 1221–1241 (2005)
8. Khati, P.: Comparative analysis of protein classification methods. Master Thesis. University of Nebraska, Lincoln (2004)
9. Cui, C., Wang, D., Shi, J.: Comparing 3-D Protein Structures Similarity by Using Fractal Features. In: 2004 IEEE Computational Systems Bioinformatics Conference (2004)
10. Daras, P., Zarpalas, D., Axenopoulos, A., Tzovaras, D., Strintzis, M.G.: Three-Dimensional Shape-Structure Comparison Method for Protein Classification. IEEE/ACM Transactions on Computational Biology and Bioinformatics 3(3), 193–207 (2006)
11. Ankerst, M., Kastenmuller, G., Kriegel, H.P., Seidl, T.: Nearest Neighbor Classification in 3D Protein Databases. In: Proc. Seventh International Conference on Intelligent Systems for Molecular Biology (ISMB 1999), Heidelberg, Germany, pp. 34–43 (1999)
12. Holm, L., Sander, C.: Protein structure comparison by alignment of distance matrices. Journal of Molecular Biology 3, 123–138 (1993)
13. Dali database, http://ekhidna.biocenter.helsinki.fi/dali/

Protein Function Prediction Based on Neighborhood Profiles

Kire Trivodaliev, Ivana Cingovska, Slobodan Kalajdziski, and Danco Davcev

Ss. Cyril and Methodius University, Faculty of Electrical Engineering and Information
Technologies, Computer Science Department, Karpos 2 BB,
1000 Skopje, Macedonia
{kiret,ivanac,skalaj,etfdav}@feit.ukim.edu.mk

Abstract. The recent advent of high throughput methods has generated large amounts of protein interaction network (PIN) data. A significant number of proteins in such networks remain uncharacterized and predicting their function remains a major challenge. A number of existing techniques assume that proteins with similar functions are topologically close in the network. Our hypothesis is that the simultaneous activity of sometimes functionally diverse functional agents comprises higher level processes in different regions of the PIN. We propose a two-phase approach. First we extract the neighborhood profile of a protein using Random Walks with Restarts. We then employ a “chi-square method”, which assigns k functions to an uncharacterized protein, with the k largest chi-square scores. We applied our method on protein physical interaction data and protein complex data, which showed the later perform better. We performed leave-one-out validation to measure the accuracy of the predictions, revealing significant improvements over previous techniques.

Keywords: Protein interaction networks, Neighbourhood extraction, Protein function prediction.

1 Introduction

The rapid development of genomics and proteomics has generated an unprecedented amount of data for multiple model organisms. As has been commonly realized, the acquisition of data is but a preliminary step, and a true challenge lies in developing effective means to analyze such data and endow them with physical or functional meaning [1]. The problem of function prediction of newly discovered genes has traditionally been approached using sequence/structure homology coupled with manual verification in the wet lab.

The first step, referred to as computational function prediction, facilitates the functional annotation by directing the experimental design to a narrow set of possible annotations for unstudied proteins.

Significant amount of data used for computational function prediction is produced by high-throughput techniques. Methods like Microarray co-expression analysis and Yeast2Hybrid experiments have allowed the construction of large interaction networks. A protein interaction network (PIN) consists of nodes representing proteins, and edges representing interactions between proteins. Such networks are stochastic as edges are weighted with the probability of interaction. There is more information in a PIN compared to sequence or structure alone. A network provides a global view of the context of each gene/protein. Hence, the next stage of computational function prediction is characterized by the use of a protein's interaction context within the network to predict its functions. A node in a PIN is annotated with one or more functional terms. Multiple and sometimes unrelated annotations can occur due to multiple active binding sites or possibly multiple stable tertiary conformations of a protein. The annotation terms are commonly based on an ontology. A major effort in this direction is the Gene Ontology (GO) project [2]. GO characterizes proteins in three major aspects: molecular function, biological process and cellular localization. Molecular functions describe activities performed by individual gene products and sometimes by a group of gene products. Biological processes organize groups of interactions into "ordered assemblies." They are easier to predict since they localize in the network. In this paper, we seek to predict the GO molecular functions for uncharacterized (target) proteins. The main idea behind our function prediction technique is that function inference using only local network analysis but without the examination of global patterns is not general enough to cover all possible annotation trends that emerge in a PIN.

According to a recent survey [3], most existing network-based function prediction methods can be classified in two groups: module assisted and direct methods. Module assisted methods detect network modules and then perform a module-wide annotation enrichment [4]. The methods in this group differ in the manner they identify modules. Some use graph clustering [5, 6] while others use hierarchical clustering based on network distance [4, 7, 8], common interactors [9] and Markov random fields [10].

Direct methods assume that neighboring proteins in the network have similar functional annotations. The Majority method [11] predicts the three prevailing annotations among the direct interactors of a target protein. This idea has later been generalized to higher levels in the network [12]. Another approach, Indirect Neighbor [13], distinguishes between direct and indirect functional associations, considering level 1 and level 2 associations. The Functional Flow method [14] simulates a network flow of annotations from annotated proteins to target ones. Karaoz et al. [15] propose an annotation technique that maximizes edges between proteins with the same function.

A common drawback of both the direct and module-assisted methods is their hypothesis that proteins with similar functions are always topologically close in the network. The direct methods are further limited to utilize information about neighbors up to a certain level. Thus, they are unable to predict the functions of proteins surrounded by unannotated interaction partners.

We hypothesize that the simultaneous activity of sometimes functionally diverse functional agents comprise higher level processes in different regions of the PIN. Our hypothesis is more general, since a clique of similar function proteins can be equivalently treated as a set of nodes that observe the same functional neighborhood. A justification for our approach is provided by Fig.1 which shows that proteins of similar function may occur at large network distances.

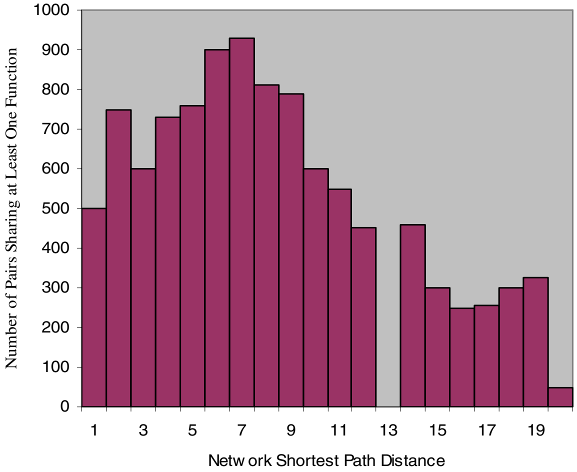


Fig. 1. Proteins sharing annotations do not always interact in the Filtered Yeast Interactome (FYI) [16]

2 Research Methods

Our approach divides function prediction into two steps: extraction of neighbourhood profile, and prediction based on the computed neighbourhood (Fig. 2). According to our hypothesis, we summarize the functional network context of a target protein in the neighbourhood extraction step. We compute the steady state distribution of a *Random Walk with Restarts (RWR)* from the protein. The steady state is then transformed into a functional profile. In the second step, we employ a *chi-sqare* method to predict the function of a target protein based on its neighbourhood profile.

2.1 Extraction of Neighbourhood Profiles

We summarize a protein's neighborhood by computing the steady state distribution of a *Random Walk with Restarts (RWR)*. We simulate the trajectory of a random walker that starts from the target protein and moves to its neighbors with a probability proportional to the weight of each connecting edge. We keep the random walker close to the original node in order to explore its local neighborhood, by allowing transitions to the original node with a probability of *c*, the restart probability.

Let *G = (V;E)* be the graph representing a protein-protein interaction network, where *V* is the set of nodes (proteins), and *E* is the set of weighted undirected edges, where the weight shows the probability of interaction (or functional association) between protein pairs. We define the proximity of a node *v* to a start node *s*, $p_s(v)$, as the steady state probability that a random walk starting at node s will end at node v.

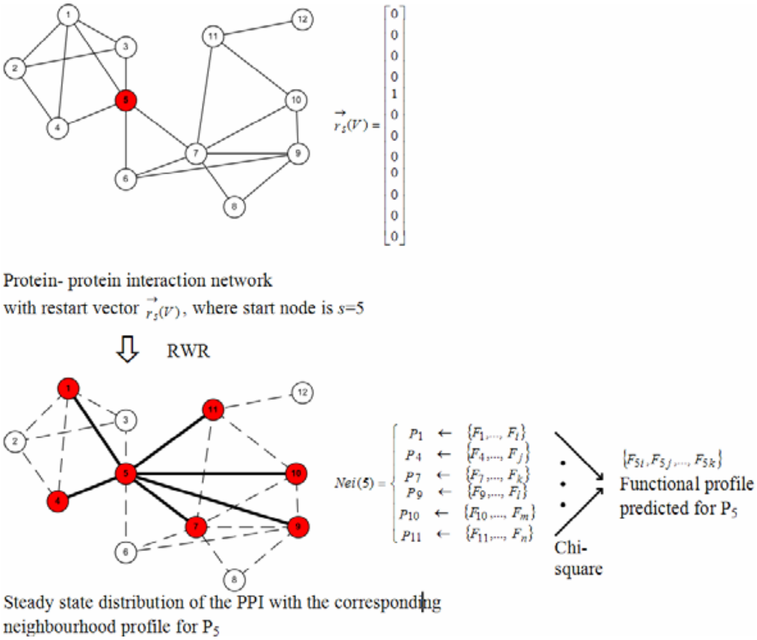

1
2
3
4
5
6
7
8
9
10
11
12
$\vec{r}_s(V) =$
$[0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0]^T$
Protein- protein interaction network
with restart vector $\vec{r}_s(V)$, where start node is s=5
RWR
$Nei(5) =$
$P_1 \leftarrow \{F_1,\dots,F_i\}$
$P_4 \leftarrow \{F_4,\dots,F_j\}$
$P_7 \leftarrow \{F_7,\dots,F_k\}$
$P_9 \leftarrow \{F_9,\dots,F_l\}$
$P_{10} \leftarrow \{F_{10},\dots,F_m\}$
$P_{11} \leftarrow \{F_{11},\dots,F_n\}$
Chi-square
$\{F_{5i}, F_{5j}, \dots, F_{5k}\}$
Functional profile predicted for P_5
Steady state distribution of the PPI with the corresponding neighbourhood profile for P_5

Fig. 2. Function prediction process: extraction of neighbourhood profile, and prediction based on the computed neighbourhood

Random walk method simulates a random walker that starts on a source node, *s* (or a set of source nodes simultaneously). At every time tick, the walker chooses randomly among the available edges (based on edge weights), or goes back to node s with probability *c*. The restart probability *c* enforces a restriction on how far we want the random walker to get away from the start node *s*. In other words, if *c* is close to 1, the affinity vector reflects the local structure around *s*, and as *c* gets close to 0, a more global view is observed.

The probability $p_s(v)^{(t)}$, describes the probability of finding the random walker at node *v* at time *t*. The steady state probability $p_s(v)$ gives a measure of proximity to node *s*, and can be computed efficiently using iterative matrix operations. Fig. 3 shows the iterative algorithm, which provably converges. The number of iterations to converge is closely related to the restart probability *c*. As *c* gets smaller the diameter of the observed neighborhood increases, thus the number of iterations to converge gets larger. The convergence check requires the L$_1$-norm between consecutive $\vec{p}_s(V)$ s to be less than a small threshold, e.g., 10^{-12}. In our experiments, for *c*=0,30 the average number of iterations to converge is around 55.

A possible interpretation of the neighborhood profile is an affinity vector of the target node to all other nodes based solely on the network structure.

```
**Input:** the interaction network $G = (V;E)$;
        a start node $s$;
        restart probability $c$;
**Output:** the proximity vector $\vec{p_s}(V)$;

Let $\vec{r_s}(V)$ be the restart vector with 0 for all its entries except a 1 for the entry denoted by node $s$;
Let **A** be the column normalized adjacency matrix defined by $E$;
Initialize $\vec{p_s}(V) := \vec{r_s}(V)$;
while ( $\vec{p_s}(V)$ has not converged):
        $\vec{p_s}(V) := (1 - c)A\vec{p_s}(V) + c\vec{r_s}(V)$;
```

Fig. 3. The iterative algorithm to compute the proximity of all the nodes in the graph to a given start node *s*

2.2 Chi-Square Method for Protein Function Prediction

The second step in our approach is predicting the annotations of a given protein P_i based on its *neighborhood profile* Nei(i). We use a method to infer protein functions based on $\chi2$-statistics. For a protein P_i, let $n_i(j)$ be the number of proteins interacting with P_i and having function F_j. Let $e_i(j)$ = #Nei(i) × π_j be the expected number of proteins in Nei(i) having function F_j , where #Nei(i) is the number of proteins in Nei(i). Define

$$S_i(j) = \frac{(n_i(j) - e_i(j))^2}{e_i(j)} \quad (1)$$

For a fixed k, they assign an unannotated protein with k functions having the top k $\chi2$-statistics.

The two steps of our approach are completely independent, different approaches can be adopted for neighboorhood extraction and classification.

3 Results

Our study assessed the reliability of two different groups of protein-protein interaction data: the protein physical interaction data and the protein complex data. The protein physical interaction data included two yeast two-hybrid data sets, by Uetz et al. [17]

and DIP [18], a collection of protein interactions from the literature and the yeast two-hybrid assays. The protein complex data included experimentally determined MIPS physical interaction data set [19], obtained by systematic purification of protein complexes and protein identification via mass spectrometry, and a set of experimentally determined protein complexes called “MIPS Complex” [19].

We separated protein complex data from protein physical interaction data because of their obvious difference: not all protein pairs in a complex interact with one another, and not all physically interacting protein pairs are in the same complex. Many protein complexes such as ribosomes and RNA Polymerases are essential for a cell, and the interactions within a complex are generally more stable and stronger and have a longer life span than most other physical interactions, while other physical interactions include other important interactions such as signal transductions.

We compare the accuracy of the techniques by performing leave one-out validation experiments. We use leave-one-out validation because many annotations in the actual network are of relatively low frequency, and thus limiting the training set. Our method is working with actual networks, containing significant number of uncharacterized proteins and hence this is a realistic measure of the accuracy. In this setup, a target protein is held out (i.e. its annotations are considered unknown) and a prediction is computed using the rest of the annotation information in the network.

The accuracy of the predictions is measured as follows. The method randomly selects an annotated protein and assumes it as unannotated. Then we predict its functions by using our method. We then compare the predictions with the annotations of the protein. We repeat the leave-one-out experiment for *K* proteins, P_i, … , P_K. Let n_i be the number of known functions for protein *Pi*, m_i be the number of *predicted* functions for protein P_i, and k_i be the overlap between the set of observed functions and the set of predicted functions. The specificity (SP) and the sensitivity (SN) can be defined as:

$$SP = \frac{\sum_i^K k_i}{\sum_i^K m_i} \quad . \tag{2}$$

$$SN = \frac{\sum_i^K k_i}{\sum_i^K n_i} \quad . \tag{3}$$

For comparison, we implement the neighbourhood counting method [11] and the $\chi 2$ method [12] for functional annotation. We choose the top 1, 2, 3, 4 and 5 functions, respectively, and assign these functions to each unannotated protein. The results of the comparison with the Uetz, DIP, MIPS physical and MIPS complex data sets are shown in Figures 4,5,6 and 7 respectively. As can be seen our method outperforms by a margin on every single data set. Figure 8 presents the assessment of the data sets in terms of their reliability when used in the process of protein function prediction. Our results confirm that the components of a protein complex can be assigned to functions that the complex carries out within a cell. The complex data sets generally perform better in function predictions than do the physical interaction data sets.

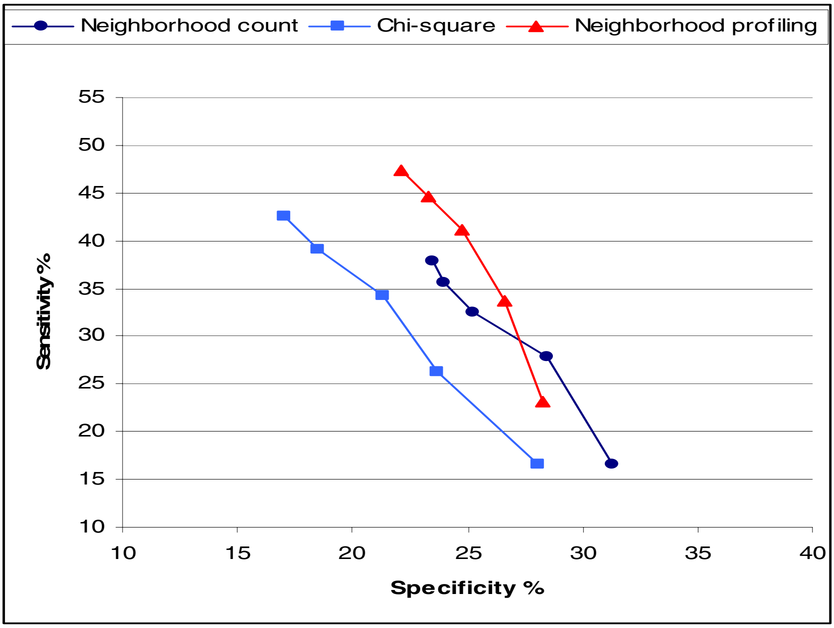


Fig. 4. Sensitivity and specificity of functional prediction using the Uetz data set

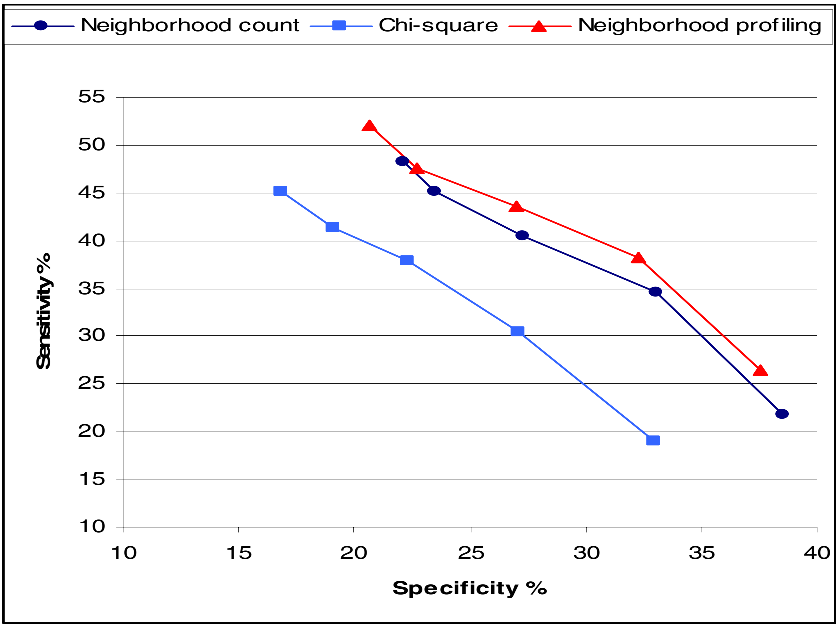


Fig. 5. Sensitivity and specificity of functional prediction using the DIP data set

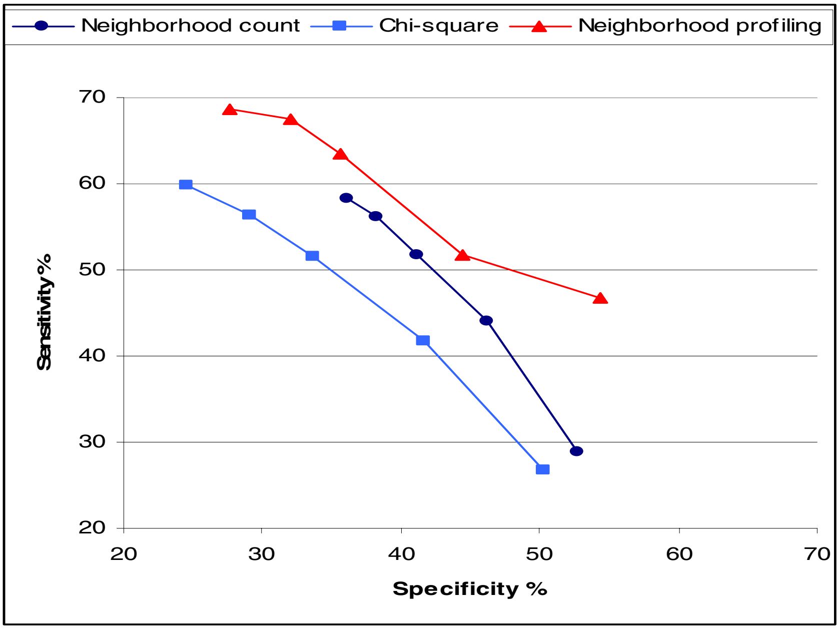


Fig. 6. Sensitivity and specificity of functional prediction using the MIPS physical data set

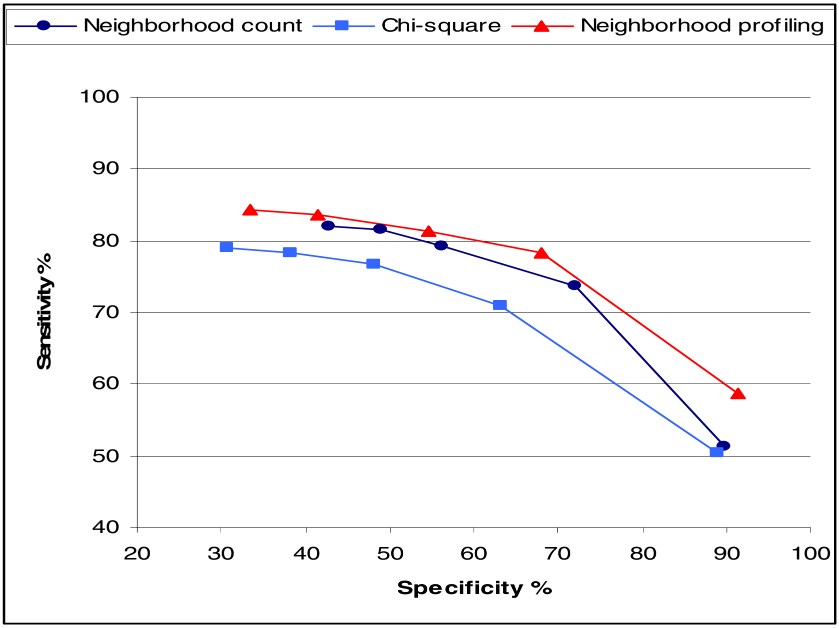


Fig. 7. Sensitivity and specificity of functional prediction using the MIPS complex data set

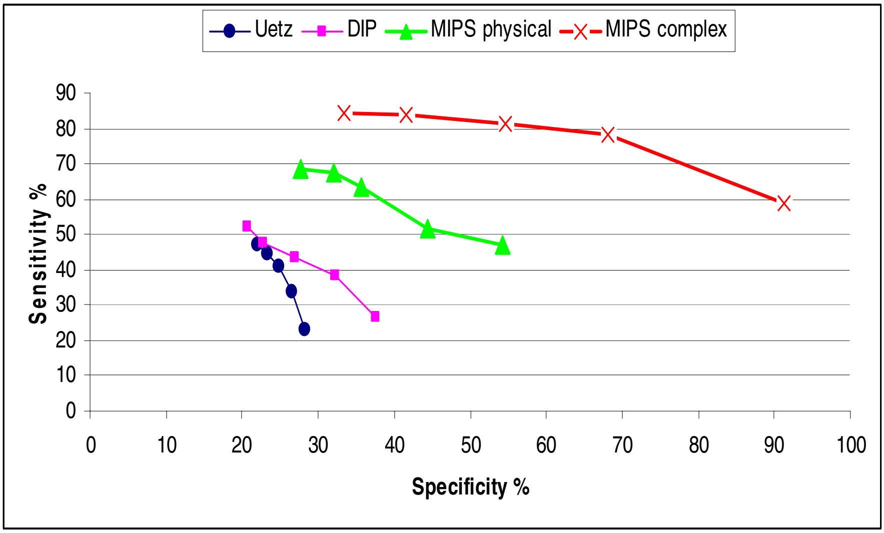


Fig. 8. Sensitivity and specificity of functional prediction for different PIN data sets

4 Conclusion

A new method for protein function prediction using protein interaction networks was presented. It is based on a general hypothesis, since a clique of similar function proteins can be equivalently treated as a set of nodes that observe the same functional neighborhood. We exploit this hypothesis by employing Random Walk with Restarts on the PIN, and extracting the neighborhood profile of an unannotated protein from which we later make the decision of assigning functions to the target by using a "chi-square" method. We validated this two-phase approach by applying it to two different groups of protein interaction data: protein physical interaction data and protein complex data. Experiments revealed that the prediction accuracy of our method outperforms existing techniques by a margin regardless of the data set used. These results are one more proof of the hypothesis that we based our method on. We also assessed the different data sets regarding their reliability on protein function prediction which showed that complex data sets generally perform better than do the physical interaction data sets.

References

1. Yu, G.X., Glass, E.M., Karonis, N.T., Maltsev, N.: Knowledge-based voting algorithm for automated protein functional annotation. PROTEINS: Structure, Function, and Bioinformatics 61, 907–917 (2005)
2. The gene ontology consortium: Gene ontology: Tool for the unification of biology. Nature Genetics 25(1), 25–29 (2000)

3. Sharan, R., Ulitsky, I., Shamir, R.: Network-based prediction of protein function. Molecular Systems Biology 3, 88 (2007)
4. Maciag, K., Altschuler, S., Slack, M., Krogan, N., Emili, A., Greenblatt, J., Maniatis, T., Wu, L.: Systems-level analyses identify extensive coupling among gene expression machines. Molecular Systems Biology 2, 2006.0003 (2006)
5. Spirin, V., Mirny, L.: Protein complexes and functional modules in molecular networks. PNAS 101, 12123–12128 (2003)
6. Dunn, R., Dudbridge, F., Sanderson, C.: The use of edge-betweenness clustering to investigate the biological function in protein interaction networks. BMC Bioinformatics 6, 39 (2005)
7. Arnau, V., Mars, S., Marin, I.: Iterative clustering analysis of protein interaction data. Bioinformatics 21(3), 364–378 (2005)
8. Brun, C., Chevenet, F., Martin, D., Wojcik, J., Guenoche, A., Jacq, B.: Functional classification of proteins for the prediction of cellular function from a protein-protein interaction network. Genome Biology 5, R6 (2003)
9. Samanta, M.P., Liang, S.: Predicting protein functions from redundancies in large-scale protein interaction networks. PNAS 100, 12579–12583 (2003)
10. Letovsky, S., Kasif, S.: Predicting protein function from protein/protein interaction data: a probabilistic approach. Bioinformatics 19, i197–i204 (2003)
11. Schwikowski, B., Uetz, P., Fields, S.: A network of protein-protein interactions in yeast. Nature Biotechnology 18, 1257–1261 (2000)
12. Hishigaki, H., Nakai, K., Ono, T., Tanigami, A., Takagi, T.: Assessment of prediction accuracy of protein function from protein–protein interaction data. Yeast 18, 523–531 (2001)
13. Chua, H., Sung, W., Wong, L.: Exploiting indirect neighbors and topological weight to predict protein function from protein-protein interactions. Bioinformatics 22(13), 1623–1630 (2006)
14. Nabieva, E., Jim, K., Agarwal, A., Chazelle, B., Singh, M.: Whole-proteome prediction of protein function via graph-theoretic analysis of interaction maps. Bioinformatics 21, i302–i310 (2005)
15. Karaoz, U., Murali, T.M., Letovsky, S., Zheng, Y., Ding, C., Cantor, C.R., Kasif, S.: Whole-genome annotation by using evidence integration in functional-linkage networks. PNAS 101, 2888–2893 (2004)
16. Han, J., Bertin, N., Hao, T., et al.: Evidence for dynamically organized modularity in the yeast protein-protein interaction network. Nature 430, 88–93 (2004)
17. Uetz, P., et al.: A comprehensive analysis of protein-protein interactions in Saccharomyces cerevisiae. Nature 403, 623–627 (2000)
18. Xenarios, I., Fernandez, E., Salwinski, L., Duan, X.J., Thompson, M.J., Marcotte, E.M., Eisenberg, D.: DIP: The Database of Interacting Proteins: 2001 update. Nucleic Acids Res. 29, 239–241 (2001)
19. Mewes, H.W., Frishman, D., Güldener, U., Mannhaupt, G., Mayer, K., Mokrejs, M., Morgenstern, B., Münsterkötter, M., Rudd, S., Weil, B.: MIPS: a database for genomes and protein sequences. Nucleic Acids Res. 30, 31–34 (2002)

Automated Structural Classification of Proteins by Using Decision Trees and Structural Protein Features

Slobodan Kalajdziski, Bojan Pepik, Ilinka Ivanovska, Georgina Mirceva,
Kire Trivodaliev, and Danco Davcev

Ss. Cyril and Methodious University, Faculty of Electrical Engineering and Information Technologies, Karpos 2 bb, 1000 Skopje, Macedonia
{skalaj,georgina,kiret,etfdav}@feit.ukim.edu.mk,
bojan.pepik@gmail.com, ilinka.ivanovska@yahoo.com

Abstract. The protein function is tightly related to classification of proteins in hierarchical levels where proteins share same or similar functions. One of the most relevant protein classification schemes is the structural classification of proteins (SCOP). The SCOP scheme has one negative drawback; due to its manual classification methods, the dynamic of classification of new proteins is much slower than the dynamic of discovering novel protein structures in the protein data bank (PDB). In this work, we propose two approaches for automated protein classification. We extract protein descriptors from the structural coordinates stored in the PDB files. Then we apply C4.5 algorithm to select the most appropriate descriptor features for protein classification based on the SCOP hierarchy. We propose novel classification approach by introducing a bottom-up classification flow, and a multi-level classification approach. The results show that these approaches are much faster than other similar algorithms with comparable accuracy.

Keywords: Structural Classification of Proteins (SCOP), C4.5 Classification, Protein function prediction.

1 Introduction

Proteins play vital structural and functional role in every cell in living organisms. They are constructed by long chains of amino acid residues folding into complex three-dimensional polypeptide chain structures. This three-dimensional representation of a residue sequence and the way this sequence folds in the 3D space are very important to understand the logic in which a function of a protein is based on. In fact, the concept of function typically acts as an umbrella term for all types of activities that a protein is involved in, be it cellular, molecular or physiological. Also, evolutionary evidence could potentially be derived from conserved protein structures existed in multiple spices. The knowledge of protein function is a crucial link in the development of new drugs, better crops, and even development of synthetic biochemicals.

Since the determining of the first 3D structure of the protein *myoglobin*, up to now, the complexity and the variety of the protein structures has increased as the number of the new determined macromolecules has. Therefore, a need for a classification of proteins is obvious, which may result in a better understanding of these complicated three-dimensional structures, their functions, and the deeper evolutionary procedures that led to their creation. In molecular biology, many classification schemes and databases (CATH [16], FSSP [15] and SCOP [2]) have been developed in order to describe the different kinds of similarity between proteins.

The Structural Classification of Proteins - SCOP database [2] describes the evolutionary relationships between proteins of known structure. It has been accepted as the most relevant and the most reliable classification hierarchy [3]. This is due to the fact that SCOP strictly builds its classification decisions based on visual observations of the structural elements of the proteins made by human experts. Therefore, this manual approach during the classification process of new structures clarifies that SCOP is completely biased towards reliable and precise protein classification. The main levels of the SCOP hierarchy are Family, Superfamily, Fold, and Class. Using the terminology of the SCOP database, two proteins that belong to the same fold share a common three-dimensional pattern with the same major secondary structure elements (SSEs) in the same arrangement with the same topological connections. In the SCOP hierarchy, folds are grouped into different classes, where a class is defined by the topographical arrangement of the secondary structures of its member proteins. Although SCOP is highly reliable and precise system, it has one negative drawback. Namely, due to its manual classification methods, the dynamic of classification of new proteins in SCOP can’t follow the dynamic of discovering novel protein structures stored in PDB (38.200 proteins classified in SCOP vs. 59.800 protein entries in PDB in August 2009). This clearly brings in front the necessity of a system which will classify proteins in a precise and reliable manner as SCOP does, but in an automated fashion.

There are various approaches for protein classification which are trying to offer efficient and completely automated protein classification. These approaches have different characteristics in terms of algorithm for determining protein similarity. Basically, the protein similarity metric used defines the complexity and the efficiency of the classification approach.

One way to determine protein similarity is to use sequence alignment algorithms like Needleman–Wunch [19], BLAST [18], PSI-BLAST [17] etc. They offer fast and efficient recognition of overlapping subsequences in two protein structures which leads to detection of closely related protein structures, but these methods cannot recognize proteins with remote homology.

Instead of sequence alignment methods, structure alignment methods like CE [5], MAMMOTH [6], DALI [7], etc. are used to detect and highlight distant homology relations between protein structures. In general these methods are very precise and efficient and they have high degree of successful mapping of existing structures in new proteins. Structure alignment methods perform one-against-all proteins comparison in order to find the most similar existing protein to a novel protein structure. Having in mind that the number of classified proteins, for example in SCOP, is ever increasing and that structure alignment methods are quite cost expensive, the speed of classification with these methods is always questioned. For example CE takes 209 days [5] to classify 11.000 novel protein structures. The bottom-up classification approaches proposed in this paper took around 6 hours to classify 9.994 proteins.

There are numerous research approaches that combine sequence and structure alignment of the proteins. SCOPmap [8] is a system that uses a pipelined architecture for the classification. SCOPmap uses four sequence alignment methods: BLAST [18], PSI-BLAST [17], RPS-BLAST and COMPASS [11] and two structure alignment methods: VAST [20] and DaliLite [10]. This pipelined approach brings to front high complexity of the classification process. FastSCOP [9] is another, more efficient system than SCOPmap which is based on 3D-BLAST [21] and MAMMOTH [6]. 3D-BLAST is structure alignment method that is used as a preprocessing filter to produce the top 10 scores. Afterwards, these top 10 results are used by MAMMOTH in order to find the most similar protein structures to the query structure. Although fastSCOP possesses high precision, the used combination of methods eventually in future, considering the ever increasing number of novel proteins, will produce increasing classification complexity.

Instead of using the alignment methods, the classification based on the mapping of the protein structure in the high-dimensional uniform descriptor space can be found as very promising. In [14] protein descriptor is formed by first producing distance matrix, which is treated as image, and local and global protein features are extracted from the image histograms. The whole descriptor dimension is 33, consisted from 24 local features and 9 global features. This protein descriptor afterwards is used for protein classification into the SCOP hierarchy based on the E-predict algorithm [14]. In [1] protein descriptor is generated solely from the protein sequence information in order to avoid complex structure comparison. The protein descriptor gives information for the number of different amino acids, the hydrophobicity, the polarity, the Van der Waals volume, the polarizability and for the secondary structures in the protein structure. With this protein descriptor, proteins are classified hierarchically into the SCOP hierarchy with Naive Bayes and boosted C4.5 methods [1].

In this work, we propose two classification processes based on generated protein descriptors in combination with C4.5 decision tree classification algorithm. As a classification scheme, the SCOP classification hierarchy is used. First we introduce the classification flow that is based on a bottom-up classification according to the SCOP hierarchy. The implemented classification logic is original and new due to the fact that according to the related work there were no protein classification approaches which use similar classification architecture. Second we adjust the multi-level modification of the C.4.5 decision tree algorithm to solve the SCOP classification. The implemented approaches introduce tremendous speed up compared with the structure alignment algorithms. They are ~816 times faster than CE [5] and ~68 times faster than MAMMOTH [6]. They are less correct than fastSCOP [9] which has accuracy of 98% for the SUPERFAMILY level compared with our 84% for bottom-up approach and 80% for multi-level approach. However, our algorithms are ~70 times faster than MAMMOTH.

In section 2 we present the classification process architecture and the used classification methods. Section 3 presents the experimental results, while the section 4 concludes the paper.

2 Classification Process Architecture and Methods

The classification process architecture has three main features. First, this architecture is based, and it uses the SCOP classification scheme. Second, it uses 3D protein descriptor [13] which transforms the protein tertiary structure into N-dimensional feature vector, and additionally gives some other protein structural features. And finally, as a classification algorithm decision trees trained with C4.5 are used.

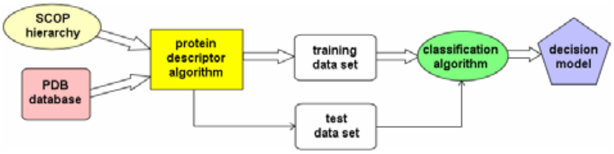


Fig. 1. Classification process architecture

Chronologically (as can be seen from the Fig. 1) the system is consisted of two phases: training phase and testing or classification phase. The training or offline data flow takes into consideration the knowledge given in the SCOP hierarchical database to build predicative classification flow for each SCOP hierarchy level. The training procedure can be divided in two general processes. The first one is the descriptor extraction (shown on Fig. 2) and data set generation process. Descriptors consisting of 450 features (416 of them describe the protein's geometry, while 34 of them give information for the primary and secondary protein structure) are generated for each protein forming a training set for the C4.5 decision tree algorithm. This descriptor relies on the geometric 3D structure of the proteins. After triangulation, normalization and voxelization of the 3D protein structures, the Spherical Trace Transform is applied to them to produce geometry - based descriptors, which are completely rotation invariant.

The second process is the process of forming and training of the decision trees for the protein classifiers. We propose two approaches in solving this classification task.

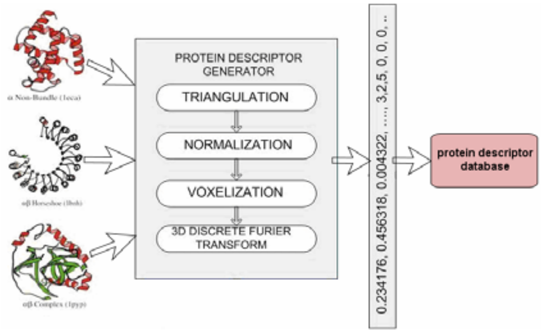


Fig. 2. Protein descriptor generation process

First one is the bottom-up classification approach, while the second one is the multi-level modification of the C4.5 algorithm. The classification logic of both classifiers is based on the fact that the SCOP classification hierarchy is tree–like hierarchy, providing only one parent node for every child node in the hierarchy. According to this fact, if we know the *domain* of the protein, then we know the upper SCOP levels (the whole hierarchy) for that protein. These two classification approaches are explained in the following subsections.

2.1 Bottom-Up Classification Approach

Basically this classification flow is by all means very much similar to the classical top – down approach, except that the starting point is changed. Instead of starting the classification from the root, it is started from the leafs.

Decision trees are built for each level of the SCOP hierarchy, providing separate trees for classification in *class*, *fold*, *superfamily*, *family* and *domain*. Also we provide additional level-specific decision trees that are trained for classification in specific

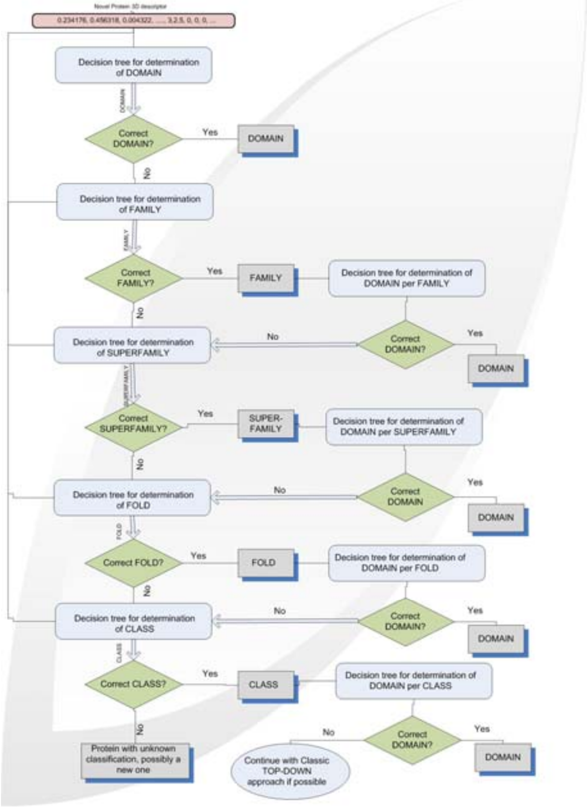

Novel Protein 3D descriptor
0.234176, 0.456318, 0.004322, ..., 3,2,5, 0, 0, 0, ...
Decision tree for determination of DOMAIN
DOMAIN
Correct DOMAIN?
Yes
DOMAIN
No
Decision tree for determination of FAMILY
FAMILY
Correct FAMILY?
Yes
FAMILY
Decision tree for determination of DOMAIN per FAMILY
No
Decision tree for determination of SUPERFAMILY
No
Correct DOMAIN?
Yes
DOMAIN
SUPERFAMILY
Correct SUPERFAMILY?
Yes
SUPER-FAMILY
Decision tree for determination of DOMAIN per SUPERFAMILY
No
Decision tree for determination of FOLD
No
Correct DOMAIN
Yes
DOMAIN
FOLD
Correct FOLD?
Yes
FOLD
Decision tree for determination of DOMAIN per FOLD
No
Decision tree for determination of CLASS
No
Correct DOMAIN?
Yes
DOMAIN
CLASS
Correct CLASS?
Yes
CLASS
Decision tree for determination of DOMAIN per CLASS
No
Protein with unknown classification, possibly a new one
No
Continue with Classic TOP-DOWN approach if possible
Correct DOMAIN?
Yes
DOMAIN

Fig. 3. Bottom-up classification approach

cases. Namely if we use *domain* specific decision tree trained for all protein instances it is obvious that this classifier will be low accurate. In cases when we know the upper level of the SCOP hierarchy of a given protein, than we can use additional decision trees trained on the subset of the protein data taking into account only the descendant proteins grouped in the lower levels of the SCOP hierarchy. In this way, there are *class* specific, *fold* specific, *superfamily* specific and *family* specific decision trees for determination of *domain*. Also, additional decision trees are built for determination of one-level up SCOP levels (*class* specific decision trees for *fold* determination etc.). These decision trees are afterwards used in the classification process of novel proteins.

As can be seen from the Fig. 3, the unknown protein is passed through the *domain* specific classifier. If the *domain* classifier correctly classifies unknown protein, then there is no need to classify the protein in the upper levels. The classification process can associate the upper hierarchy labels from the background knowledge extracted from the SCOP hierarchy. If the *domain* classifier incorrectly classifies the protein, than the protein is being preceded to the classifier into the higher level of the hierarchy, in this case the decision tree for *family* determination. If the classified *family* is correct, than the rest of the levels of the hierarchy for the protein are known, except the *domain*, which remains unknown. To correct this, we precede the protein to the level-specific decision tree for determination of *domain*, trained only with instances from the predicted *family*. In this way we lower the false positive hits in the classification flow if we use separate decision trees for separate SCOP levels. If the classified *family* is mistaken, the classifier continues one level up into the hierarchy, in this case it continues with predicting the *superfamily* level of the new protein. This step is also the next step if the *family* specific decision tree mistakes the protein *domain*. Otherwise the classification is finished. The process of protein classification continues with the backward recursion explained in the previous paragraph until it reaches the top level of the SCOP hierarchy, the *class* level. If there is no success in recognizing the correct protein *domain*, then the protein is announced as a protein with unknown SCOP label, possibly a candidate for a new label in the SCOP hierarchy.

2.2 Multi-level Modification of the C4.5 Algorithm

Traditional *single-label* classification is concerned with learning from a set of examples that are associated with a single label l from a set of disjoint labels L, $|L|>1$. If $|L|=2$, then the learning problem is called a *binary* classification problem, while if $|L|>2$, then it is called a *multi-class* classification problem. The problem of classifying proteins in SCOP hierarchy is a standard *multi-class* or *multi-level* classification problem.

For the purposes of the protein classification, we have used and readapted the modification of the C4.5 algorithm [22] for multi-label data. In order to automate the SCOP classification we need to: (1) have information about the hierarchy of classes, (2) calculate the entropy, and (3) check the membership of the new protein in the existing hierarchy. We have provided flat text file with the SCOP hierarchy labels organized in tree-like manner. The modification of the C4.5 algorithm for multi-label data is made by changing the entropy calculation:

$$entropy(S) = -\sum_{i=1}^{N}\left(p(c_i)\log p(c_i) + q(c_i)\log q(c_i)\right) \quad (1)$$

where $p(c_i)$ = relative frequency of class c_i and $q(c_i) = 1-p(c_i)$. Also there is allowed multiple labels in the leaves of the tree.

By applying the multi-label C4.5 classification the end result is one decision tree that can classify new protein structures in branch of the SCOP hierarchy at once.

3 Experimental Results

All of the experiments were conducted on a PC with a 2.2 GHz Intel Core 2 CPU and 2GB RAM. The SCOP hierarchy from the last two versions SCOP1.71 and SCOP 1.73 were downloaded and integrated into Oracle 10g database. The protein descriptor generation was made in C++. The bottom-up classification approach was implemented in C#.NET by using the C4.5 decision trees generated by WEKA data mining toolkit. The multi-level modification of the C4.5 algorithm was implemented in C#.NET.

In the training phase, 73.642 out of 75.930 classified protein chains from SCOP v1.71 were used for training the decision trees for both classification strategies. For the bottom-up approach, instead of having one decision tree for each level of the SCOP hierarchy, ensemble of decision trees is used for each level of the SCOP hierarchy. The idea for ensemble of trees came as a result of the huge memory requirements of the approach with one tree per level. The number of trees per ensemble in one level and number of output classes per tree in each level are presented in Table 1.

Table 1. Number of trees per ensemble in one level and number of output classes per tree in each level

Level	Classes per tree	Number of trees per level
CLASS	11	1
FOLD	256	5
SUPERFAMILY	169	12
FAMILY	341	11
DOMAIN	659	14

In the test phase, 3.576 protein chains from the SCOP v1.73 were taken. We have selected only those proteins that were not previously classified in the SCOP v1.71, and the *domain* of the selected protein from the SCOP v1.73 must be present in the SCOP v1.71 hierarchy.

The classification results for our proposed classification approaches, bottom-up classification approach and multi-level modified C4.5 approach are shown on the Table 2 and Table 3 respectively.

Table 2. Results of the classification with the bottom-up approach

Level	Correctly classified	Incorrectly classified	Accuracy
CLASS	3154	422	88,2%
FOLD	3027	549	84,65%
SUPERFAMILY	2993	583	83,69%
FAMILY	2886	690	80,7%
DOMAIN	2839	737	79,39%

Table 3. Results of the classification with the multi-level modification of C4.5

Level	Correctly classified	Incorrectly classified	Accuracy
CLASS	3050	526	85,29%
FOLD	2957	619	82,69%
SUPERFAMILY	2864	712	80,09%
FAMILY	2839	737	79,38%
DOMAIN	2821	755	78,88%

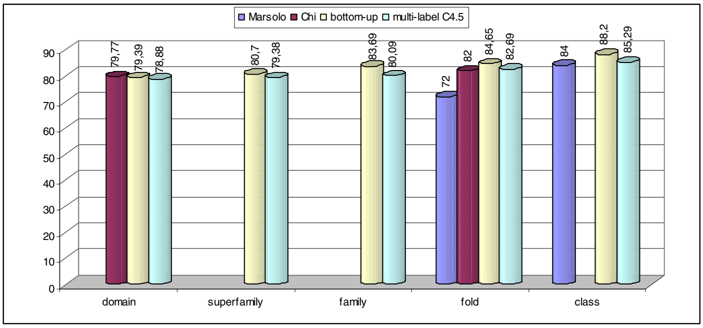


Fig. 4. Classification accuracy comparison of our approaches and the approaches presented in [14] and [1]

We have compared our results with the protein classification approaches given in [1] and [14]. The dataset used in [1] is based on SCOP v1.67 and consists of 311 training proteins taken from 27 most populated SCOP folds with no more than 35% sequence similarity between any two proteins. This approach presents only classifications by *class* and *fold* levels of the SCOP hierarchy. The dataset used in [14] is based on SCOP v1.69, and the test proteins are taken among two consequence SCOP versions. This approach provides only classification results for the *fold* level of the SCOP hierarchy. On Fig.4 we show the comparison results between classification results given in [1], [14], and our bottom-up approach and multi-level modification of C4.5 algorithm. As can be seen from the obtained results, the precision of the classification is satisfying. In classifying *fold* in [14] the precision is 92% for SCOP v1.69 when E-predict is used as a classification algorithm, but if C4.5 decision tree is used as classifier the precision for *fold* prediction is 82%. From this point of view our classifiers have comparable accuracy and produce classification results for the whole SCOP hierarchy, not by partial levels. It should be mentioned here, that fastSCOP [9] predicts the protein *superfamily* with 98% accuracy compared with 84% accuracy in our bottom-up classifier, but our bottom-up classifier classifies proteins nearly ~70 times faster then fastSCOP thus providing much higher efficiency.

Also we have conducted the speed performance testing (as shown on the Fig. 5) of our classification approaches compared with the CE [5] and MAMMOTH [6] approaches. We have randomly selected around 10.000 proteins and passed them to all four systems. The bottom-up classification lasted 6 hours on a Intel Core 2 Duo machine with 2 GB RAM, while multi-level C4.5 modification lasted around 11,5 hours (~1,95 times slower) on the same machine. The MAMMOTH results are obtained on Intel Pentium 2.8 GHz machine, while the CE results are obtained on Sun Ultra Sparc II machine.

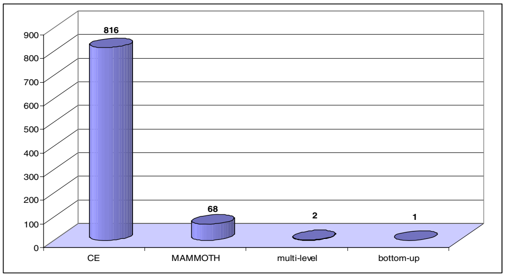


Fig. 5. Speed comparison of protein classification by using different classification approaches

4 Conclusion

The objective of this paper was to make empirical tests of the usefulness and the contribution of the different protein features to the decision tree classification precision and the usefulness of the bottom-up classification approach and multi-level modification of the C4.5 algorithm. It is evident that these approaches which use protein descriptor and decision tree algorithms for classification introduces high level of efficiency which can be concluded from the time taken to classify unknown protein and the percent of correctly classified proteins. The multi-level modification of C4.5 does not find as many rules as would be found by learning all the levels individually. This is to be expected, as the criteria for choosing nodes in the decision tree are slightly different, and a different amount of information is available.

The provided comparison with some other relevant works proves the satisfying results obtained with this work. In the future we plan to extend the classification in order to solve the problem of classification of proteins in novel SCOP branches.

References

1. Marsolo, K., Parthasarathy, S., Ding, C.: A Multi-Level Approach to SCOP Fold Recognition. In: IEEE Symposium on Bioinformatics and Bioeng., pp. 57–64 (2005)
2. Murzin, A.G., Brenner, S.E., Hubbard, T., Chothia, C.: Scop: a structural classification of proteins database for the investigation of sequences and structures. Journal of Molecular Biology 247, 536–540 (1995)

3. Camoğlu, O., Can, T., Singh, A.K., Wang, Y.F.: Decision tree based information integration for automated protein classification. Journal of Bioinformatics and Computational Biology 3(3), 717–724 (2005)
4. Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N., Bourne, P.E.: The Protein Data Bank. Nucleic Acids Res. 28, 235–242 (2000)
5. Shindyalov, H.N., Bourne, P.E.: Protein structure alignment by incremental combinatorial extension (CE) of the optimal path. Protein Engineering 9, 739–747 (1998)
6. Ortiz, A.R., Strauss, C.E., Olmea, O.: Mammoth: An automated method for model comparison. Protein Science 11, 2606–2621 (2002)
7. Holm, L., Sander, C.: Protein structure comparison by alignment of distance matrices. Journal of Molecular Biology 233, 123–138 (1993)
8. Cheek, S., Qi, Y., Krishna, S.S., Kinch, L.N., Grishin, N.V.: SCOPmap: Automated assignment of protein structures to evolutionary superfamilies. BMC Bioinformatics 5, 197–221 (2004)
9. Tung, C.H., Yang, J.M.: FastSCOP: a fast web server for recognizing protein structural domains and SCOP superfamilies. Nucleic Acids Res. 35, W438–W443 (2007)
10. Holm, L., Sander, C.: Dali: a network tool for protein structure comparison. Trends in Biochemical Science 20, 478–480 (1995)
11. Sadreyev, R., Grishin, N.: COMPASS: a tool for comparison of multiple protein alignments with assessment of statistical significance. J. Mol. Biol. 326, 317–336 (2003)
12. Yang, J.M., Tung, C.H.: Protein structure database search and evolutionary classification. Nucleic Acids Research 34, 3646–3659 (2006)
13. Kalajdziski, S., Mirceva, G., Trivodaliev, K., Davcev, D.: Protein Classification by Matching 3D Structures. In: Frontiers in the Convergence of Bioscience and Information Technologies 2007, Jeju Island, Korea, pp. 147–152 (2007)
14. Chi, P.H.: Efficient protein tertiary structure retrievals and classifications using content based comparison algorithms. PhD thesis, University of Missouri-Columbia (2007)
15. Holm, L., Sander, C.: The FSSP Database: Fold Classification Based on Structure-Structure Alignment of Proteins. Nucleic Acids Research 24, 206–210 (1996)
16. Orengo, C.A., Michie, A.D., Jones, D.T., Swindells, M.B., Thornton, J.M.: CATH - A hierarchic classif. of protein domain structures. Structure 5(8), 1093–1108 (1997)
17. Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W., Lipman, D.J.: Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Research 25(17), 3389–3402 (1997)
18. Altschul, S.F., Gish, W., Miller, W., Myers, E.W., Lipman, D.J.: Basic local alignment search tool. Journal of Molecular Biology 215(3), 403–410 (1990)
19. Needleman, S.B., Wunsch, C.D.: A general method applicable to the search for similarities in the amino acid sequence of two proteins. J. Mol. Bio. 48(3), 443–453 (1970)
20. Madej, T., Gibrat, J.F., Bryant, S.H.: Threading a database of protein cores. Proteins 23, 356–369 (1995)
21. Tung, C.H., Huang, J.W., Yang, J.M.: Kappa-alpha plot derived structural alphabet and BLOSUM-like substitution matrix for rapid search of protein structure database. Genome Biology 8(3), 31–46 (2007)
22. Clare, A.: Machine learning and data mining for yeast functional genomics. PhD thesis, University of Wales Aberystwyth (2003)

Wireless Sensor Networks Localization Methods: Multidimensional Scaling vs. Semidefinite Programming Approach

Biljana Stojkoska, Ilinka Ivanoska, and Danco Davcev

Faculty of Electrical Engineering and Information Technologies – Skopje, Karpoš II bb,
1000 Skopje, Macedonia
biles@feit.ukim.edu.mk, ilinka_iv@yahooo.com,
etfdav@feit.ukim.edu.mk

Abstract. With the recent development of technology, wireless sensor networks are becoming an important part of many applications such as health and medical applications, military applications, agriculture monitoring, home and office applications, environmental monitoring, etc. Knowing the location of a sensor is important, but GPS receivers and sophisticated sensors are too expensive and require processing power. Therefore, the localization wireless sensor network problem is a growing field of interest. The aim of this paper is to give a comparison of wireless sensor network localization methods, and therefore, multidimensional scaling and semidefinite programming are chosen for this research. Multidimensional scaling is a simple mathematical technique widely-discussed that solves the wireless sensor networks localization problem. In contrast, semidefinite programming is a relatively new field of optimization with a growing use, although being more complex. In this paper, using extensive simulations, a detailed overview of these two approaches is given, regarding different network topologies, various network parameters and performance issues. The performances of both techniques are highly satisfactory and estimation errors are minimal.

Keywords: Wireless Sensor Networks, Semidefinite programming, multi-dimensional scaling, localization techniques.

1 Introduction

New technologies bring new possibilities, however, in the same time new questions are being opened. The area of wireless sensor networks solves a great amount of new problems. A wireless sensor network (WSN) is a network consisting of distributed sensor devices that cooperatively monitor physical or environmental conditions at different locations. The development of wireless sensor networks was originally motivated by military applications. However, wireless sensor networks are now used in many industrial and civilian application areas, including industrial process monitoring and control, machine health monitoring, environment and habitat monitoring, healthcare applications and traffic control. Today, wireless sensor networks has become a key technology for different types of smart environments, and the aim is to enable the application of wireless sensor networks for a wide range of industrial problems. Wireless networks are of particular importance when a large number of sensor nodes have to be deployed.

A fundamental problem in wireless sensor networks is localization i.e. the determination of the geographical locations of sensors. Localization is a challenge when dealing with wireless sensor nodes, and a problem which has been studied for many years [1]. Nodes can be equipped with a Global Positioning System (GPS), but this is a costly solution in terms of money and power consumption. The localization issue is important where there is an uncertainty about some positioning. If the sensor network is used for monitoring the temperature in a remote forest, nodes may be deployed from an airplane and the precise location of most sensors may be unknown. An effective localization algorithm can then use all the available information from the nodes to compute all the positions.

Most existing localization algorithms were designed to work well in wireless sensor networks. The performance of localization algorithms depend on critical sensor network parameters, such as the radio range, the network topology i.e. the density of nodes, the anchor-to-node ratio, and it is important that the solution gives adequate performance over a range of reasonable parameter values.

In this paper we give an overview of two completely different localization approaches: Multidimensional scaling and Semidefinite programming. We present analysis and simulations of the algorithms, demonstrating the accuracy compared to each other, regarding different sensor network parameters.

The Multidimensional scaling approach is an algorithm using connectivity information for computing the nodes' localization with the help of some linear transformations [2]. The MDS-MAP algorithm first uses connectivity to roughly estimate the distance between each pair of nodes, then, multidimensional scaling (MDS) is used to find possible node locations that fit the estimations, and finally, it is optimized by using the anchors positions [3]. In section 2 we describe the classical MDS approach used in the simulations.

Section 3 describes the Semidefinite programming (SDP) relaxation based method for the position estimation problem in sensor networks [4][5]. The basic idea behind the technique is to convert the nonconvex quadratic distance constraints into convex constraints by introducing a relaxation to remove the quadratic term in the formulation. The solving of the connection convex constraint is by using techniques of linear programming.

In Section 4 we show results on both simulated algorithms with a discussion and a comparison of the two proposed methods and finally, Section 5 concludes the paper.

2 Multidimensional Scaling

First we will give a mathematical model of the wireless sensor network localization problem. In all the localization approaches the network is modeled by a graph $G=(V,E)$, where V is the set of nodes some with known positions in the Euclidean space $R^{\dim}$ and E is the set of edges defined by the network topology (connectivity). In one case a set of weights $\{d_{ij}:(i,j)\in E\}$ on the graph's edges is given, representing the (estimated) distances between the corresponding nodes. The problem is then to place all nodes in such a way that the Euclidean distance between every pair of nodes v and w, where $(v,w)\in E$, equals d_{vw}. In the other case, d_{ij} are not given a special value. It is only assumed that $d_{ij} < R$, where R is the range of the transmitter of a wireless sensor node. To describe the positions of the nodes of the network, we form a corresponding matrix and to store the available distance information we define the matrix $D = \left\{ d_{ij}\ i, j = 1, 2, ..., n \right\}$.

For the Multidimensional scaling approach we consider the node localization problem with defining the network as an undirected graph with vertices *V* and edges *E*. The vertices correspond to the nodes, of which zero or more may be special nodes, which we call anchors, whose positions are already known. We assume that all the nodes being considered in the positioning problem form a connected graph, i.e., there is a path between every pair of nodes.

We focus on classical MDS in this paper. Classical MDS is the simplest case of MDS: the proximities of objects are treated as distances in a Euclidean space. The goal of MDS is to find a configuration of points in a multidimensional space such that the inter-point distances are related to the provided proximities by some transformation (e.g., a linear transformation).

Let p_{ij} refer to the proximity measure between objects i and j. The Euclidean distance between two points $X_i = (x_{i1}; x_{i2}; x_{i3}...x_{im})$ and $X_j = (x_{j1}; x_{j2}; x_{j3}...x_{jm})$ in an m - dimensional space is

$$d_{ij} = \sqrt{\sum_{k=1}^{m} (x_{ik} - x_{jk})^2} \qquad (1)$$

The Euclidean distances are related to the proximities by a transformation $d_{ij} = f(p_{ij})$. In the classical MDS, a linear transformation model is assumed, i.e. $d_{ij} = a + bp_{ij}$. The distances D are determined so that they are as close to the proximities P as possible. In that way, we define $I(P) = D + E$, where $I(P)$ is a linear transformation of the proximities, and E is a matrix of errors. Since D is a function of the coordinates X, the goal of classical MDS is to calculate X such that the sum of squares of E is minimized, subject to suitable normalization of X.

In classical MDS, P is shifted to the center and coordinates X can be computed from the double centered P through singular value decomposition. For an $n \times n$ P matrix for n points and m dimensions of each point, it can be shown that

$$-\frac{1}{2}({p_{ij}}^2 - \frac{1}{n}\sum_{i=1}^{n} {p_{ij}}^2 - \frac{1}{n}\sum_{i=1}^{n} {p_{ij}}^2 + \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} {p_{ij}}^2) \qquad (2)$$

$$= \sum_{k=1}^{m} x_{ik} x_{jk}$$

The double centered matrix on the left hand side (B) is symmetric. Having now calculated B and performing singular value decomposition on B gives $B = VAV$. The coordinate matrix becomes $X = VA^{1/2}$.

Retaining the first r largest eigenvalues and eigenvectors $(r < m)$ leads to a solution in lower dimension. This implies that the summation over k runs from 1 to r instead of m. This is the best low rank approximation in the least-squares sense. For example, for a 2D network, we take the first 2 largest eigenvalues and eigenvectors to construct the best 2D approximation.

2.1 MDS-MAP Algorithm

MDS-MAP is a localization method based on multidimensional scaling [3]. The MDS-MAP algorithm consists of three steps:

- Compute the shortest distances between all pairs of nodes in the region. The computed distances are used for building the distance matrix for MDS.
- Apply classical MDS to the distance matrix, retaining the first 2 eigenvalues and eigenvectors to construct a 2D relative map.
- Given sufficient anchor nodes (3 or more), transform the relative map to an absolute map based on the absolute positions of anchors.

In the first step, we assign distances to the edges in the connectivity graph. When the distance of a pair of neighbor nodes is known, the value of the corresponding edge is the measured distance. When we only have connectivity information, a simple approximation is to assign value 1 to all edges. Then a classical all-pairs shortest-path algorithm, such as Dijkstra's algorithm, can be applied. In the second step, classical MDS is applied directly to the distance matrix. The core of classical MDS is singular value decomposition. The result of MDS is a relative map that gives a location for each node. Although these locations may be accurate relative to one another, the entire map will be arbitrarily rotated relative to the true node positions. In the third step, the relative map is transformed through linear transformations, which include scaling, rotation, and reflection. The goal is to minimize the sum of squares of the errors between the true positions of the anchors and their transformed positions in the MDS map.

3 Semidefinite Programming

Semidefinite programming is the other approach we are going to present in this paper. It is a new optimization algorithm that uses techniques of linear programming.

It will be helpful to first introduce some mathematical notations to describe this technique. The trace of a given matrix A, denoted by $t_r(A)$ is the sum of the entries on the main diagonal of A. A symmetrical matrix is called semidefinite if all its eigenvalues are nonnegative and is represented by $A \succeq 0$.

Suppose two nodes x_1 and x_2 are within radio range R of each other, the proximity constraint can be represented as a convex second order cone constraint of the form $\left\| x_1 - x_2 \right\|_2 \le R$, and this can be formulated as a matrix linear inequality.

$$\begin{pmatrix} I_2 R & x_1 - x_2 \\ (x_1 - x_2)^T & R \end{pmatrix} \succeq 0 \tag{3}$$

The mathematical model of the localization problem can be described as follows. There are $n-m$ distinct sensor points in $R^{\dim}$ whose locations are to be determined, and other m fixed points (called the anchor points) whose locations are known. The known nodes are indicated by a and the unknown nodes by $\hat{x}$, so that $X = \left[\hat{x}_1, ...\hat{x}_{n-m}, a_1, ..., a_m \right]$. All $(i, j) \in E$ where $i < j$ and if j is an anchor are denoted by N_a , and all $(i, j) \in E$, where $i < j$ are unknown is denoted by N_x . The following constraints must be satisfied:

$$\begin{aligned} \| a_k - \hat{x}_j \|^2 = d^2{}_{kj} \quad & \forall (k, j) \in N_a \\ \| \hat{x}_i - \hat{x}_j \|^2 = d^2{}_{ij} \quad & \forall (i, j) \in N_x \end{aligned} \tag{4}$$

We consider the case when the node distances are known, therefore, let $X = \left[\hat{x}_1; ...; \hat{x}_{n-m} \right]^T$ be the matrix in $R^{\dim \times (n-m)}$ that needs to be determined. Define $e_{ij} \in R^{n-m}$ with 1 on i-th position and with -1 on j-th position, and everywhere else zeros. If $I_{\dim}$ is the identity matrix, the constraints can be written:

$$\begin{aligned} (a_k; e_j)^T \left[I_{\dim}; X^T \right] [I_{\dim}; X] (a_k; e_j) = d_{kj}^2 \quad & \forall (k, j) \in N_a \\ e_{ij}^T X^T X e_{ij} = d_{ij}^2 \quad & \forall (i, j) \in N_x \end{aligned} \tag{5}$$

We now need to find a symmetric matrix $Y \in R^{\dim \times \dim}$ and X that satisfy the following constraints:

$$\begin{aligned} (a_k; e_j)^T \begin{pmatrix} I_{\dim} & X \\ X^T & Y \end{pmatrix} (a_k; e_j) = d_{kj}^2 \quad & \forall (k, j) \in N_a \\ e_{ij}^T Y e_{ij} = d_{ij}^2 \quad & \forall (i, j) \in N_x \\ Y = X^T X & \end{aligned} \tag{6}$$

This is the SDP formulation of the problem of wireless sensor networks localization. In [4] a relaxation of this method is proposed that we will use in our simulations. The constraint $Y = X^T X$ is relaxed with $Y \succeq X^T X$. We can write this condition as follows

$$Z = \begin{pmatrix} I_{\dim} & X \\ X^T & Y \end{pmatrix} \succeq 0 \qquad (7)$$

In this way the SDP problem can be written as $\min \; 0$, such that

$$Z_{1:\dim,1:\dim} = I_{\dim} \qquad (8)$$

$$(0;e_{ij})(0;e_{ij})^T \bullet Z = d_{ij}{}^2 \quad \forall(i,j) \in N_x$$

$$(a_k;e_j)(a_k;e_j)^T \bullet Z = d_{kj}{}^2 \quad \forall(k,j) \in N_a$$

$$Z \succeq 0$$

Where $A \bullet B = t_r(AB)$ and 0 the zero vector of the corresponding dimension. When we have a solution to this problem, we can then easily extract the solution for the positions of the unknown nodes, since they are then defined by X and are a part of Z. Practically this is solved by a SDP solver such as SeDuMi which we used in our simulations.

4 Simulation Results

In our experiments, we ran MDS-MAP and SDP algorithms on various topologies of networks in Matlab. Three different network topologies were considered: (1) random topology with a uniform distribution within a square area, (2) square grid topology with some placement errors, and (3) on a hexagonal grid topology with some placement errors.

For the SDP approach the computational results presented here were generated using the interior-point algorithm SDP solvers SeDuMi [6] with their interfaces to Matlab.

In a square grid, with a placement error, a random value drawing from a normal distribution N(0;1) is added to the node's original grid position. The placement error in a hexagonal grid topology is defined similarly.

The data points represent averages over 20 rounds in networks containing 64 nodes. The anchor nodes are selected randomly and the number of anchor nodes varies between 4 and 10 in each simulation. The connectivity (average number of neighbors) is controlled by specifying radio range R. Nodes are placed in a square area with size of rxr (r=0.5).

4.1 Random Network Topology

In the case when the network has a random topology, 64 nodes are placed randomly in rxr square area (r=0.5). Figure 1 shows an example of this random placement and the results in the SDP approach are given. The radio range here is 0.15r, which leads to an average connectivity of 14.625, and the number of anchor nodes is 6. Figure 2 shows a comparison of MDS-MAP and SDP estimation errors in a random network topology with 64 nodes placed in a square area with size of rxr (r=0.5). The radio ranges (R) used are 0.15r, 0.18r, 0.2r and 0.25r, which lead to an average connectivity level of 13.31, 18.21, 21.55 and 29.75 respectively. The number of the anchor nodes used is 4, 5, 6 and 10.

Figure 2 shows a better performance of the estimated errors in the SDP approach in comparison with the MDS-MAP algorithm. The estimation errors of the SDP

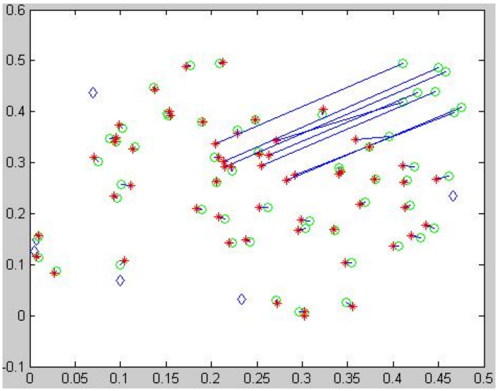


Fig. 1. SDP simulation of a random network topology with 64 nodes placed in a square area 0.5x0.5 and 6 anchors, average connectivity 14.625

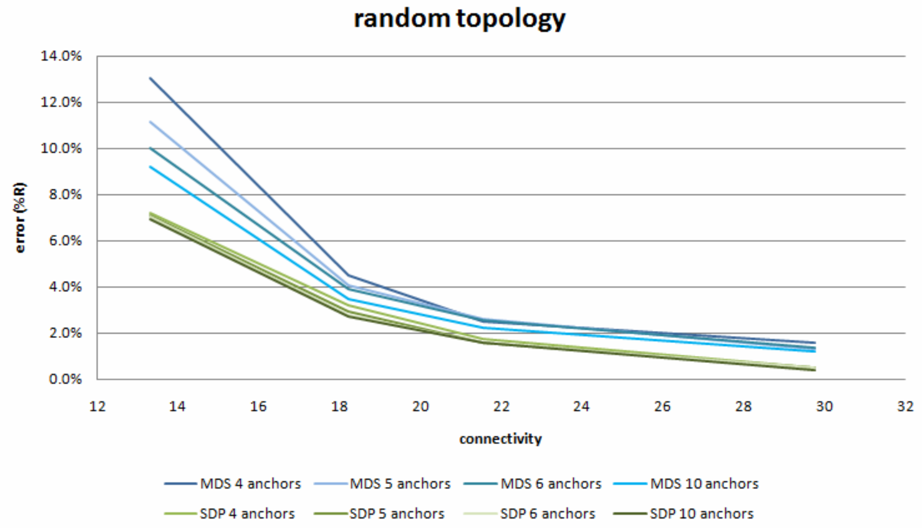


Fig. 2. Comparison of MDS-MAP and SDP estimation errors in a random network topology with 64 nodes placed in a square area with size of 0.5x0.5

approach are almost 2 times smaller than the MDS estimation errors. For example, MDS gives an estimation error of 0.1302R with connectivity level of 13.3125, and in contrast SDP estimation error is 0.0721R (in a case with 4 anchors), and moreover, for connectivity level of 29 and more SDP estimation errors are less than 0.01R. Obviously when the connectivity level rises, estimation errors are getting smaller even by half for connectivity level less than 18. An interesting result is that the number of anchor nodes does not effect much on the SDP estimation errors.

4.2 Square Grid Network Topology

In the case when the network has a square grid topology, we assume that the sensor nodes are deployed according to a regular structure. Actually, nodes are placed in the neighborhood of the vertices due to random placement error. 64 nodes are placed on a rxr (r=0.5) grid, with a unit edge distance r/8. This type of network topology is shown in figure 3. The results in the SDP approach are given here. The radio range is 0.15r, which leads to an average connectivity of 13.055, and the number of anchor nodes is 6. Figure 4 shows a comparison of MDS-MAP and SDP estimation errors in a square grid network topology with 64 nodes placed in a square area with size of rxr (r=0.5). The radio ranges (R) used are 0.15r, 0.18r, 0.2r and 0.25r, which lead to a average connectivity level of 12.97, 18.07, 21.19 and 30.14 respectively.

Our results show that MDS and SDP obtain much better results on the grid layout than on the random layout for the same connectivity level. Estimation errors are lowered by half with this regular topology in comparison with the random topology. SDP outperforms MDS in the same way as in the random placement.

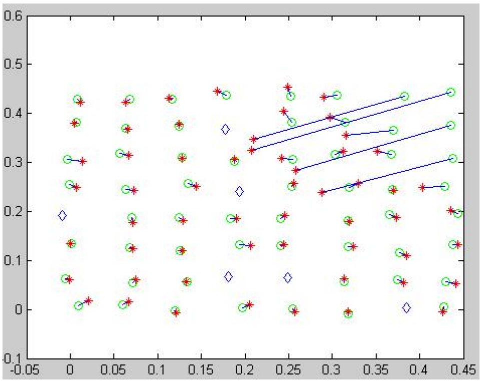


Fig. 3. SDP simulation of a square grid network topology with 64 nodes placed in a square area 0.5x0.5 and 6 anchors, average connectivity 13.375

4.3 Hexagonal Grid Network Topology

The case when the network has a hexagonal grid topology, is similar with the square grid topology. Sensor nodes are placed on the vertices of a hexagonal grid with a random placement error as in figure 5. Figure 6 shows a comparison of MDS-MAP

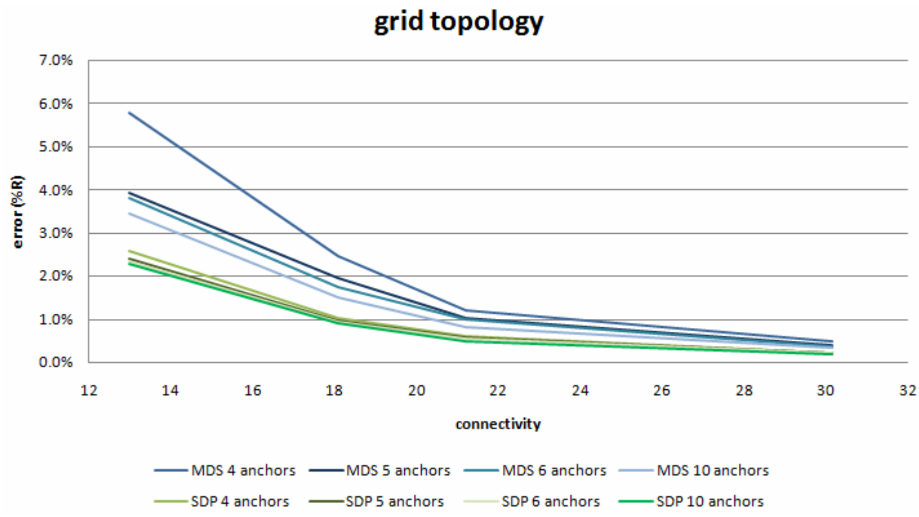


Fig. 4. Comparison of MDS-MAP and SDP estimation errors in a square grid network topology with 64 nodes placed in a square area with size of 0.5x0.5

and SDP estimation errors in a hexagonal grid network topology with 64 nodes placed in a square area with size of rxr (r=0.5). The radio ranges (R) used are 0.15r, 0.18r, 0.2r and 0.25r, which lead to a average connectivity level of 14.94, 20.35, 24.075 and 33.45 respectively. The simulation results here are similar with the square grid case. SDP estimation errors are lower than MDS estimation errors, but the improvement in the estimation errors with the SDP approach is not as stressed as in the random layout.

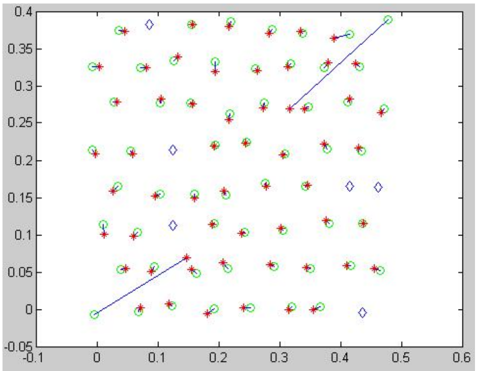


Fig. 5. SDP simulation of a hexagonal grid network topology with 64 nodes placed in a square area 0.5x0.5 and 6 anchors, average connectivity 14.925

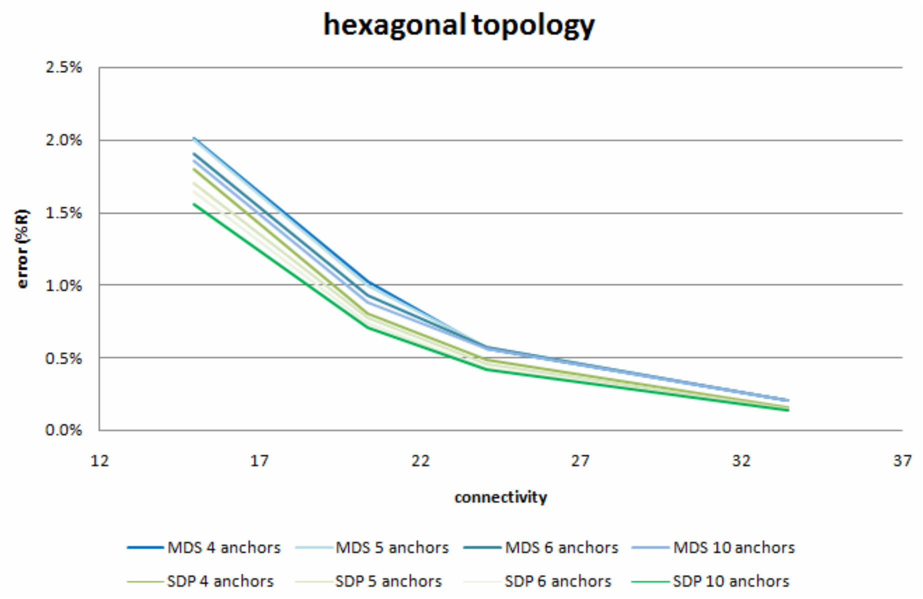


Fig. 6. Comparison of MDS-MAP and SDP estimation errors in a hexagonal grid network topology with 64 nodes placed in a square area with size of 0.5x0.5

5 Conclusions

In the vast field of research related to wireless sensor networks, our focus has been on the problem of localization, one of the major challenges in the design of ad hoc networks. Our goal was to present two approaches that are of a rising interest: Multidimensional scaling with the MDS-MAP algorithm and localization with Semidefinite programming. They both work well, with a small amount of connectivity information about the network, however the Semidefinite programming approach with known distance network information outstands with its results in comparison to the Multidimensional Scaling approach. In conclusion, SDP as an approach is better than MDS for small sized networks (as used in our simulations) especially with random topologies. However, the time consuming factor is not considered. SDP as a more complex algorithm is slower than MDS, and for larger network sizes it will be very difficult to get any results. Although some research has been done concerning network topologies, some other irregular topologies should be considered in future with different network sizes. Furthermore, some hybrid algorithms which combine the advantages of this two approaches (greater performance with SDP and speed with MDS) should be developed.

References

1. Rudafshani, M., Datta, S.: Localization in wireless sensor networks. In: 6th international conference on Information processing in sensor networks, pp. 51–60 (2007)
2. Ji, X., Zha, H.: Sensor positioning in wireless ad-hoc sensor networks using multidimensional scaling. In: 23rd Annual Joint Conference of the IEEE Computer and Communications Societies (INFOCOM), pp. 2652–2661 (2004)
3. Shang, Y., Ruml, W., Zhang, Y., Fromherz, M.: Localization from Mere Connectivity. In: 4th ACM international symposium on Mobile and Ad-Hoc Networking & Computing, pp. 201–212. ACM Press, New York (2003)
4. Biswas, P., Ye, Y.: Semidefinite programming for ad hoc wireless sensor network localization. In: Third international symposium on Information processing in sensor networks, pp. 46–54. ACM Press, New York (2004)
5. Biswas, P., Lian, T.C., Wang, T.C., Ye, Y.: Semidefinite programming based algorithms for sensor network localization. ACM Transactions on Sensor Networks (TOSN) 2(2), 188–220 (2006)
6. `http://sedumi.mcmaster.ca`

On the Complexity of the Greedy Construction of Linear Error-Correcting Codes

Dejan Spasov and Marjan Gusev

Institute of Informatics, Faculty of Natural Science, Ss. Cyril and Methodius University,
Skopje, Macedonia
{dejan,marjan}@ii.edu.mk

Abstract. Greedy algorithms are one of the oldest known methods for code construction. They are simple to define and easy to implement, but require exponential running time. Codes obtained with greedy construction have very good encoding parameters; hence, the idea of finding faster algorithms for code generation seems natural. We start with an overview of the greedy algorithms and propose some improvements. Then, we study the code parameters of long greedy codes in attempt to produce stronger estimates. It is well known that greedy-code parameters give raise to the Gilbert-Varshamov bound; improving this bound is fundamental problem in coding theory.

Keywords: Linear codes, Greedy Codes, Lexicodes, Gilbert-Varshamov Bound, Greedy Algorithms.

1 Introduction

Given an n-dimensional vector space F_q^n over some finite field F_q, a *code* C is any subset of M elements. Let $d(x,y)$ denotes the *Hamming distance* between two vectors x and y, i.e. the number of coordinates in which they differ; then we can define *minimum distance* d of a code as $d = \min(d(x,y))$, $\forall x, y \in C$. We write (n, M, d) to denote a code of M elements and minimum distance d over F_q^m.

The main focus in this paper is on the *linear codes*. The $M = q^k$ codewords of a linear code form a k-dimensional subspace in F_q^n. We write $[n,k,d]$ to denote a linear code of dimension k and minimum distance d. A linear code C has a $k \times n$ generator matrix G and $(n-k)\times n$ parity check matrix H, such that $HG^T = 0$. Throughout the paper, we will assume that the generator and parity check matrices are in standard form $G = [I \quad A]$, and $H = [-A^T \quad I]$.

We use $wt(x) \in \mathrm{N}$ to denote the Hamming weight of the vector x, i.e. the number of nonzero positions of x, $wt(x) = dist(x,0)$. In addition, δ will denote the relative distance of the code $\delta = d/n$, and R will be the code rate $R = \log_q(M)/n$. A string of n ones, $11\ldots1$, will be written as 1^n, and the concatenation of two strings a and b will be represented with $(a\,|\,b)$.

In estimating the complexity of an algorithm, we adopt Random Access Machine (RAM) as a computational model. The time complexity is measured as the number of basic (sequential) steps needed for instance of the algorithm to end. It is considered that RAM has unlimited memory with instant access. Thus the space complexity is simply the number of registers used by an instance of the algorithm.

In Section 2, we study the complexity of the greedy algorithms for code generation. In 2.3, we will introduce an algorithm that we call the Jenkins' construction. We will show that this algorithm has better time complexity than the similar constructions (2.1 and 2.2). In 2.4, we will explain the Lexicographic construction- a greedy approach that originally was introduced in [1]. In this paper we will present faster and less memory demanding version of the algorithm, generalized over arbitrary alphabet $GF(q)$. Our contribution to the algorithm is underlined with lemma 1 and theorem 1. Using them we avoid using long coset leaders and we work with their coset weights instead.

It is obvious that due to the exponential nature of the greedy algorithm it is impossible to generate long codes in an acceptable time. Thus, in section 3, we want to estimate the code parameters of the long greedy codes. In 3.1, we develop counting mechanisms (theorems 2, 3, and 4) that give better estimate on code parameters than the well-known Gilbert-Varshamov bound. In 3.2, we use theorem 3 to improve some published results (see (12) and (13)). Some parts in section 3 have already been published in [14]; we present them again for the reason of completeness.

2 Greedy Algorithms for Code Construction

Fundamental problem in coding theory is how to find *optimal codes*. The code (n,M,d) is optimal if it has maximal number of codewords M for a given n and d. In general, finding an optimal code is considered to be a difficult problem. Trivial way to do this is by super-exponential search over all possible orderings of the field F_q^n. For small fields ($q \le 9$, $n \le 256$), there exist tables of best known (some of them optimal) codes [7], but for larger spaces optimal-code parameters can be estimated with the Gilbert-Varshamov bound and its asymptotical variant.

It is well-known fact that a simple greedy search produces a code with parameters δ and R that follow the Gilbert-Varshamov bound. In binary case, unproven conjecture is that this is the best known method for code construction. A particularly mysterious truth is that almost all random linear codes meet the asymptotic Gilbert-Varshamov bound.

2.1 The Gilbert's Construction

In general, Gilbert's Construction produces a nonlinear (n,M,d) code. Given the length n and the minimum distance d, the algorithm searches over the entire space

F_q^n and greedily adds to C the first vector x, such that $d(x,c) \geq d, \forall c \in C$. The time complexity of the algorithm is $O\left(nq^{(1+R)n}\right)$, because there are q^n vectors to be checked against a set of at most M codewords, $M = q^{Rn}$.

Important to notice is that the space needed to store all M codewords is nq^{Rn}. This is one of the reasons why nonlinear codes are not very popular for practical purposes.

2.2 The Varshamov's Construction

Given the codimension $m = n - k$ and the distance d, the Varshamov's algorithm searches over F_q^m and produces the parity check matrix H of a linear code. The algorithm greedily adds to H the first vector x that is not linear combination of $d-2$ or less columns of H. It can be verified that with this construction every combination of $d-1$ columns of H is linearly independent; hence the resulting code is linear with minimum distance d.

The time complexity of the algorithm is $O\left(n^2 q^{n(1-R+H(\delta))}\right)$, because each of the q^{n-k} vectors of F_q^m must be compared with at most $q^{nH(\delta)}$ linear combinations, where $H(\delta)$ is the entropy function, and $q^{nH(\delta)}$ represents all $d-2$ or less combinations from an n-element set, when $n \to \infty$. The space complexity of the Varshamov's construction is proportional with the dimensions of the parity check matrix H.

2.3 The Jenkins' Construction

The idea in this construction is to build the generator matrix of a systematic code $G = [I \quad A]$. For each $x \in F_q^m$ we form the vector $c_{k+1} = (10\ldots0 \mid x)$, $c \in F_q^n$. Then we check if all linear combinations of rows of the matrix $G_{k+1} = \begin{bmatrix} G \\ c_{k+1} \end{bmatrix}$ have Hamming weight at least d. There are two important facts about the systematic codes that help to reduce the number and size of the row checks: every linear combination of d or more rows of G_{k+1} has Hamming weight $\geq d$, and

$$wt\left(c_{i_1} + c_{i_2} + \cdots + c_{i_N}\right) = N + wt\left(a_{i_1} + a_{i_2} + \cdots + a_{i_N}\right)$$

where each c_{i_j} is a row of G_{k+1}, and a_{i_j} are the parity bits of c_{i_j} that belong to the A-matrix.

The worst-case running time of the algorithm is $O\left(n^2 q^{n(1-R+RH(\delta/R))}\right)$, because q^{n-k} vectors must be checked in $q^{nRH(\delta/R)}$ linear combinations, while the space complexity is proportional with the size of the A-matrix. We named this algorithm Jenkins' construction because, to our knowledge, it was first implemented by B. Jenkins in [2], by using divide-and-conquer strategy in such a manner that first it checks all linear combinations of one vector, then all linear combinations of two vectors, and so on.

2.4 The Lexicographic Construction

Any linear code C, with parity check matrix H, partitions the entire space F_q^n into q^{n-k} disjoint sets of q^k elements called *cosets*. Two vectors $x, y \in F_q^n$ belong to the same coset if and only if $x - y \in C$. Each coset has two special vectors: a unique *syndrome* $s \in F_q^{n-k}$ and a *coset leader* $e(s) \in F_q^n$. The coset leader $e(s)$ is the minimum weight vector in the coset. The syndrome s is obtained with $s = H \cdot x^T$, where x is any vector from the coset. Usually, all q^{n-k} pairs $(e(s), s)$ are stored in a table of size nq^{n-k} called *standard array*. The standard array is used to describe minimum distance decoding principle called *syndrome decoding*. In syndrome decoding, the error pattern $e \in F_q^n$ that scrambled the transmitted message $y \in F_q^n$ is considered to be the coset leader $e(s)$ associated with the syndrome $s = H \cdot y^T$. The weight of the heaviest coset leader in the array is called *covering radius* ρ of the code.

In [1], A. Trachtenberg published an algorithm that uses the standard array of a binary $[n, k, d]$ code C to build a new $[n_{k+1}, k+1, d]$ code in a greedy fashion. The algorithm, named Lexicographic Construction, runs in time $\mathrm{O}(n^2 2^{n-k})$ and space $\mathrm{O}(n 2^{n-k})$. In [4], D. Spasov improved the space complexity of the Lexicographic Construction to $\Theta(2^{n-k})$, by demonstrating that the algorithm can work with the weights of the coset leaders $w(s) \in \mathrm{N}$ instead of the coset leaders $e(s)$. In addition, in [4], the time complexity was reduced to $\mathrm{O}(n 2^{n-k})$ by making the greedy choice less time consuming. In this section, we will describe the Lexicographic Construction, as implemented in [4], but generalized over $GF(q)$.

Given the code $[n, k, d]$, let assume that the pairs syndrome-coset weight $(s, w(s))$ are stored in a table. The Lexicographic Construction is iterative algorithm that can be described in three steps. In the first step the covering radius of the code is found, $\rho = \max(w(s)), \forall s \in F_q^m$. In the second step, a vector from the maximum weight cosets is chosen, and a new basis codeword is formed. In the third step the table $(s_{k+1}, w(s_{k+1}))$ for the new $[n_{k+1}, k+1, d]$ code is computed from $(s, w(s))$.

Even though, for the second step, it seems that we can not obtain a coset member from $(s, w(s))$, the following lemma shows how to get one for systematic codes:

Lemma 1. Vectors l and $(0^k \mid s)$ belong to the same coset if and only if $Hl^T = s$.

Proof. $H \cdot (0^k \mid s)^T = s$ ■

Hence, in the second step, the algorithm chooses the syndrome s of a maximum weight coset, $w(s)=\rho$, and attaches to it k zeroes, $\left(0^k \mid s\right)$. Then the new basis codeword c_{k+1} is constructed by adding $d-\rho$ ones:

$$c_{k+1}=\left(1^{d-\rho} \mid 0^k \mid s\right) \tag{1}$$

A proof that in binary case the code $\left[n_{k+1},k+1,d\right]$ will have minimum distance d can be found in [1]. This proof can be generalized for any $GF(q)$.

The third step is an effective mechanism to build the table $\left(s_{k+1},w(s_{k+1})\right)$ from the existing $\left(s,w(s)\right)$. First, note that any vector s_{k+1} that belongs to $\mathbf{F}_q^{n_{k+1}-k-1}$ is a syndrome of a coset. This syndrome s_{k+1} is considered to be a concatenation of two vectors $s_{k+1}=\left(v \mid s\right)$, such that v and $s, s\in \mathbf{F}_q^{n-k}$. Second, we need the following definition:

Definition 1. Given two syndromes s_k and s, the *syndrome companion set* of s_k with respect to s is the set:

$$\left\{y_i \mid y_i=\mathrm{s}_\mathrm{k}+i\cdot s,\ \ i\in \mathrm{F}_\mathrm{q}\right\} \tag{2}$$

There are q^{n-k-1} disjoint syndrome companion sets and each syndrome belongs to only one companion set. Next, we can proceed to the main result:

Theorem 1. Given ρ, c_{k+1}, and $\left(s_k,w(s_k)\right)$. The table $\left(s_{k+1},w(s_{k+1})\right)$ associated with $\left[n_{k+1},k+1,d\right]$ can be efficiently constructed by carrying out the following minimization for each syndrome $s_{k+1}=\left(v \mid s_k\right)$

$$w(s_{k+1})=\min_{\substack{i=0..q-1\\ y_i=s_k+i\cdot s}}\left\{wt\left(v+i^{d-\rho}\right)+w(y_i)\right\}\ . \tag{3}$$

Proof Outline. Any coset from $\mathbf{F}_q^{n_{k+1}-k-1}$ can be seen as concatenation of q companion cosets from $\mathbf{F}_q^{n-k}$. In a simple case, the new coset leader is chosen to be the minimum weight coset leader among the companion cosets. Companion cosets are found by finding the syndrome companion set (2) for each syndrome. ■

The Lexicographic Construction starts with the repetition code and iterates as long as there are available memory resources. The space complexity of the algorithm is $O\left(\log(n)2^{\lceil \lg(q)\cdot\rceil(1-R)n}\right)$. In practice, we can speak about implementation only for the case $d=const$, thus the size of the registers needed for storing $w(s_k)$ can be considered constant and the space complexity becomes $\Theta\left(2^{\lceil \lg(q)\cdot\rceil(1-R)n}\right)$.

For reference purposes, theorem 2.2 published in [5] is a generalization of theorem 3 from [1] for q-ary alphabet. Our Theorem 1, not only shows that the same conclusion can be derived for coset weights $w(s)$, but also shows how to find the coset companions. In contrast, the method used for finding coset companions in [1] is binary search.

3 Estimate of the Parameters of the Greedy Codes: The Gilbert-Varshamov Bound

Let H is the parity check matrix of some binary code $[n-1,k-1,d]$. Let $H(n-k,d-2)$ denotes the set of all unique $(n-k)$-tuples that are linear combination of $(d-2)$ columns of H. Then a code with parameters $[n,k,d]$ does exist provided

$$|H(n-k,d-2)| \leq 2^{n-k}-2 \tag{4}$$

The existence of $[n,k,d]$ lower-bounds the existence of the optimal code for given n and d. Let $V(n,d)$ denotes the number of all possible d or less combinations from an n-element set

$$V(n,d)=\sum_{i=0}^{d}\binom{n}{d}. \tag{5}$$

Then, since $|H(n-k,d-2)|$ cannot be larger than $V(n-1,d-2)$ we obtain a simple combinatorial estimate of (4)

$$V(n-1,d-2) \leq 2^{n-k}-2 \tag{6}$$

known as the Gilbert-Varshamov (GV) bound. One of the most challenging problems in coding theory is how to improve the GV bound, especially for infinite code length n. In binary case, so far, only one asymptotic improvement is known [8], and a few non-asymptotic ones (see [10], [12], and [13]).

3.1 Some Results on the Gilbert-Varshamov Bound

We start with a simple improvement of the GV bound for the case when d is even number:

Theorem 2. Let the minimum distance d is even number. Then the code $[n,k,d]$ does exist provided

$$V(n-2,d-3) \leq 2^{n-k-1}-2\,. \tag{7}$$

Proof. Construct the code $[n,k,2t]$ from the code $[n-1,k,2t-1]$ by adding overall parity check. Use (6) to find $[n-1,k,2t-1]$. ■

It can be shown that the right-hand of (7) is always smaller than the right-hand of (6) by a factor n/d. Despite the simplicity of theorem 2, we have not found any publication that mentions it. Theorem 3 was presented in [14]; however later we discovered that it was already published in similar fashion in [13]:

Theorem 3. [13] The code $[n,k,d]$ can be extended to a code with parameters $[n+l+1,k+l,d]$ provided that

$$\sum_{\substack{i=1 \\ i=i+2}}^{\min(l,d-2)} \binom{l}{i} V(n,d-2-i) \leq 2^{n-k} . \tag{8}$$

Interesting to note is that the existence of $[n,k,d]$ can be supported with the Varshamov bound (6), or by using (8) recursively. The recursion ends with the repetition code.

Proof Outline. We build the parity check matrix H_{m+1} recursively from H_m, $H_{m+1} = [H_m \quad L_m]$. In order to estimate $|H(n-k,d-2)|$, we count only those linear combinations that include odd number of vectors from L_m. The parameter l is the number of columns of the matrix L_m. ■

In absence of stronger evidence, simulations suggest that theorem 3 non-asymptotically improves the GV bound when $\delta = const$.

If we compare code parameters that are solution of (8) with the parameters obtained from running the greedy algorithm [4], we will notice a considerable gap. This implies that many linear combinations from $H(n-k,d-2)$ are counted multiple times. Theorem 4 attempts to improve this over-counting:

Theorem 4. The code $[n,k,d]$ can be extended to a code with parameters $[n+l+1,k+l,d]$ provided that

$$\begin{aligned} &\sum_{\substack{i=1 \\ i=i+2}}^{\min(l,d')} \binom{l}{i} V(n,d'-i) - \\ &-\frac{1}{2} \sum_{\substack{t=2 \\ t=t+2}}^{\min(l,d')} \binom{l}{t} \sum_{\substack{i=1 \\ i=i+2}}^{t} \binom{t}{i} \sum_{j=t-i}^{d'-i} \binom{d'}{j} V(n-d',d'-\max(i+j,t-i+d'-j)) \leq 2^{n-k} \end{aligned} . \tag{9}$$

where $d' = d-2$.

Proof Outline. We will show an example of the case when a vector is counted twice with (8). Let $H_{m+1} = [H_m \quad L_m]$. Then for any two columns $l_1, l_2 \in L_m$, there exist $d-2$ columns in H_m such that $l_1 + l_2 + h_1 + h_2 + \cdots + h_{d-2} = 0$. If we transfer some vectors on the right-hand side we obtain

$$l_1 + h_1 + h_2 + \cdots + h_i = l_2 + h_{i+1} + h_{i+2} + \cdots + h_{d-2} .$$

Hence, we observe that the vector $l_1 + h_1 + h_2 + \cdots + h_i$ is counted twice. Theorem 4 is simply a generalization of this observation that includes all linear combinations of even number of vectors from L_m. ■

For Hamming codes $(d=3)$, there is no difference between theorem 3 and 4; however, for $d=5$ inequality (9) becomes slightly better, namely

$$l\left(1+n+\binom{n}{2}\right)-3\binom{l}{2}+\binom{l}{3}\leq 2^{n-k-1}. \tag{10}$$

3.2 Comparison with Prior Work

To the best of our knowledge, theorem 3 was first published in [13] in order to estimate the parameters of the greedy lexicodes. However, in [13] it is not mentioned that $[n,k,d]$ needs not to be a greedy code, and that (8) can be used recursively, first to guarantee existence of $[n,k,d]$, then the existence of $[n+l+1,k+l,d]$.

Elia [10] reported the following result: Let the code $[n-2,k-1,d]$ does exist. Then the code $[n,k,d]$ does exist too, provided

$$V(n-2,d-3)\leq 2^{n-k-1}. \tag{11}$$

If we restrict l to be at most 1 then (8) is precisely Elia's result. Moreover, letting $l\leq 2$ we obtain improvement of (11); namely, assuming prior existence of the code $[n-3,k-2,d]$, the code $[n,k,d]$ does exist if the following holds true

$$V(n-3,d-3)\leq 2^{n-k-2}. \tag{12}$$

Even though (7) and (11) are the same inequality, they are used in a different context. In [10], inequality (11) is used only after the existence of $[n,k,d]$ is secured. In theorem 2, prior existence of $[n,k,d]$ is not required, but (7) is restricted only for codes with even minimum distance.

Barg, Guritman, and Simonis [12] reported the following remark: The code $[n,k,d]$ with covering radius $\rho\leq d-2$ can be extended to $[n+d-\rho-1,k+1,d]$. In this context, if the covering radius of $[n,k,d]$ is strictly less than $d-2$, then (8) guarantees existence of the trivial lengthening $[n+1,k,d]$. However if we have prior knowledge of the covering radius, we can modify (8) so that we obtain at least the same result as in [12]. For example, similar to (11), we can extend *remark 13* from [12], i.e. if

$$V(n-1,\alpha)\leq 2^{n-k-1}. \tag{13}$$

for some $\alpha\leq d-1$, then any $[n,k,d]$ code can be extended to an $[n+d-\rho,k+2,d]$ code. For $\alpha=d-3$ this reduces to (12).

Jiang and Vardy have developed a graph-theoretic approach to asymptotically improve the GV bound for nonlinear codes [8], [9]. They were able to show that the code (n,M,d) does exist provided

$$c\frac{V(n,d-1)}{n}\leq 2^{n-\lceil\log_2 M\rceil}. \tag{14}$$

where the constant c is at least $1/2+o(1)$, as reported in [9]. How does (8) compares with (14)? So far, we were unable to prove that the left-hand of (8) can be smaller by a factor n. Hence, one may assume that (14) guarantees existence of a code with better parameters than (8). However, in general inequality (14) guarantees existence of a non-linear code, while (8) pertains to the linear codes. Gaborit and Zemor [11] proved that some linear double circulant codes follow (14), but only for code rates of $1/2$. If a linear code is proved to comply with (14), then (8) and (14) will complement each other. Namely, Jiang and Vardy reported that (14) improves the GV bound when the relative distance δ is constant. On the other hand, (8) improves the GV bound even when δ approaches to zero.

4 Conclusion

In section 2, we have introduced four exponential-time greedy algorithms. In binary case these algorithms remain asymptotically the best known method for code construction. An open problem is to find polynomial-time construction that meets the greedy-code parameters, or to prove non-existence of such an algorithm.

The exponent in the growth rate of an exponential algorithm is the key factor that determines the running time. Our goal was to find algorithms with smaller exponents in the worst-case running time; though improving the worst-case running time not necessarily guarantees faster algorithm. More important is the average-case running time. In the case of the Lexicographic Construction the worst-case equals the average-case. However, in the case of the Jenkins' construction, we leave the average-case complexity as open problem. On the other hand, the best-case complexity of the Gilbert's construction is $\mathrm{O}(nq^k)$. Comparing this best-case with the Lexicographic construction's worst-case, we concluded that not only the Lexicographic Construction has better space complexity than the Gilbert's construction, but also it is faster for, at least, code rates $R \geq 1/2$.

In general, finding a faster algorithm is a difficult task, since a faster algorithm will have to check only a fraction of the q^{n-k} codeword candidates or only a fraction of the $\binom{k}{d}$ row checks. The solution that we propose is combination of the Lexicographic construction and the Jenkins' construction. First, as long as there are available memory resources, run the Lexicographic construction. Then, after the entire memory is used, continue with the Jenkins' construction, while the table $(s, w(s))$ is still kept in memory for reducing the number of row checks.

In section 3, we tried to count only once as many linear combinations as possible from of the parity check matrix H. The complex theorem 4 has the best possible estimate on $\left|H(n-k, d-2)\right|$. However, even for $d=5$, there is a big difference between the estimated results (11) and simulated results [4]. The obvious conclusion is that there are still many combinations that are counted multiple times, but we believe that asymptotical improvements similar to [8] may exist for linear codes.

References

1. Trachtenberg, A.: Designing Lexicographic Codes with a Given Trellis Complexity. IEEE Trans. Information Theory 48(1), 89–100 (2001)
2. Jenkins, B.: Tables of Binary Lexicodes, `http://www.burtleburtle.net/bob/math/lexicode.html`
3. Barg, A.: Complexity Issues in Coding Theory. Handbook of Coding Theory. Elsevier Science, Amsterdam (1998)
4. Spasov, D.: Implementing the Lexicographic Construction, `http://nislab.bu.edu/nislab/projects/lexicode/index.html`
5. O'Brien, K., Fitzpatrick, P.: Covering radius construction codes with minimum distance at most 8 are normal, `http://www.bcri.ucc.ie/BCRI_01.pdf`
6. Vardy, A.: Algorithmic Complexity in Coding Theory and the Minimum Distance Problem. In: STOC (1997)
7. Grassl, M.: Bounds on the minimum distance of linear codes and quantum codes, `http://www.codetables.de` (accessed on 2009-09-02)
8. Jiang, T., Vardy, A.: Asymptotic improvement of the Gilbert-Varshamov bound on the size of binary codes. IEEE Trans. Inform. Theory 50, 1655–1664 (2004)
9. Vu, V., Wu, L.: Improving the Gilbert-Varshamov bound for q-ary codes. IEEE Trans. Inform. Theory 51, 3200–3208 (2005)
10. Elia, M.: Some results on the existence of binary linear codes. IEEE Trans. Inform. Theory 29, 933–934 (1983)
11. Gaborit, P., Zemor, G.: Asymptotic improvement of the Gilbert-Varshamov bound for binary linear codes. IEEE Trans. Inform. Theory 54, 3865–3872 (2008)
12. Barg, A., Guritman, S., Simonis, J.: Strengthening the Gilbert-Varshamov Bound. Lin. Alg. Appl. 307, 119–129 (2000)
13. O'Brien, K., Fitzpatrick, P.: Improving the Varshamov bound by counting components in the Varshamov graph. Designs, Codes, and Cryptography 39(3) (2006)
14. Spasov, D., Gusev, M.: Some notes on the binary Gilbert-Varshamov bound. In: Sixth International Workshop on Optimal Codes and Related Topics, Varna, Bulgaria (2009)

Vulnerability Assessment of Complex Networks Based on Optimal Flow Measurements under Intentional Node and Edge Attacks

Igor Mishkovski[1], Risto Kojchev[2], Dimitar Trajanov[2], and Ljupco Kocarev[2,3]

[1] Politecnico di Torino, Turin, Italy
igor.mishkovski@polito.it
[2] Faculty of electrical engineering and information technologies, Skopje, Macedonia
rkojcev@gmail.com, mite@feit.ukim.edu.mk,
lkocarev@feit.ukim.edu.mk
[3] Macedonian Academy of Sciences and Arts, Skopje, Macedonia
lkocarev@manu.edu.mk

Abstract. In this paper we assess the vulnerability of different synthetic complex networks by measuring the traffic performance in presence of intentional nodes and edge attacks. We choose which nodes or edges would be attacked by using several centrality measures, such as: degree, eigenvector and betweenness centrality. In order to obtain some information about the vulnerability of the four different complex networks (random, small world, scale-free and random geometric) we analyze the throughput of these networks when the nodes or the edges are attacked using some of the above mentioned strategies. When attack happens, the bandwidth is reallocated among the flows, which affects the traffic utility. One of the obtained results shows that the scale-free network gives the best flow performance and then comes random networks, small world, and the poorest performance is given by the random geometric networks. This changes dramatically after removing some of the nodes (or edges), giving the biggest performance drop to random and scale-free networks and smallest to random geometric and small world networks.

Keywords: Vulnerability, NUM, complex networks, attack strategies, measurements, bandwidth allocation.

1 Introduction

In today's everyday life we are surrounded with complex systems. These complex systems can be represented as networks with a certain number of nodes joined together by edges. Commonly cited examples include social networks, technological networks, information networks, biological networks, communication networks, neural networks, ecological networks and other either naturally occurring or man-made occurring networks. The topology of these complex networks is one aspect that might help understand in details the surrounding complex systems and its exploration started with the graph theory introduced by Erdős and Rényi [1]. Erdős and Rényi introduced random models in order to model the real complex systems and to capture some of the main characteristics of the real complex systems. However, these models could not give a clear picture of the topology of complex systems and there was an increasing need of new more realistic models. Watts and Strogatz found out that many real world networks exhibit what is called the small world property, i.e. most vertices can be reached from the others through a small number of edges, like in social networks. After the introduction of the Watts and Strogatz's model, Barabási and Albert showed that the structure and the dynamics of the network are strongly affected by nodes with a great number of connections [2]. It was found that many real complex networks have a power-law distribution of a node's degree and by that they are in fact scale-free networks. Additionally, many of the systems are strongly clustered with a big number of short paths between the nodes, i.e. they obey the small world property. Another contribution that helped understand the underlying topology of some real complex system, such as ad hoc networks, is made by Penrose introducing the random geometric graphs and their properties [3].

The above mentioned models helped in understanding the dynamic processes that might occur in the network. Epidemic spreading [4,5], nodes' protection so that the network can resist certain attacks or failures [6], gossip [7] or the process or spreading influence in the network [8], synchronization among nodes [9], cascading failures [10] are some examples of dynamic behaviors of complex networks.

Recently, the primary interest in complex networks is the flow properties of the transport entities. In the complex systems there are many types of flows, such as: traffic flows, information flows, energy flows, chemical flows, idea flows, etc. In particular, the most interesting aspect is how the networks structure affects the flow properties, like traffic congestion [11]. In addition to this, many researchers have studied how attacks or failures of nodes affect the traffic performance in the network [12]. This is a present problem in the real-world networks like the power grids, the Internet, telephone networks and transportation networks. In [13] authors study the robustness to random and intentional node attacks. In this study when a node is attacked, the flows which go through the node have to reconfigure their paths which may affect the loads on the other nodes and may start a sequence of overload failures. Their results show that scale-free networks are highly robust to random node failures but fragile to intentional node attacks, while the random graphs are robust under both node attacks. In their results, the flow rates are assumed to be fixed even after the reconfiguration of flow paths. In [14] authors study the effect of random and intentional attacks on the traffic performance in the Internet. They define some indicators to measure the traffic performance and show how they are affected. In [15] authors analyze the total throughput of ad hoc networks with different network interaction models at communication level, such as: random, small world, scale-free, geographic, full mesh and star models. Their results show that the full-mesh network has highest throughput, while scale-free and star networks show lowest throughput.

In this paper we are assessing the vulnerability of complex networks based on optimal flow measurements under intentional node and edge attacks. We are using four models of complex networks as underlying networks: random, small world, scale-free and geometric model. On these models we calculate the optimal bandwidth allocation solution for a given flow scenario. Then we are measuring the vulnerability of the network by using different strategies for node and edge removal and calculating the reduction of the total flow under the given scenario, network model and attack strategy.

Therefore, the main goal of this work is to measure and analyze the vulnerability of different complex networks under different node and edge strategies by measuring the total flow in the network.

The rest of the paper is organized as follows. In Section 2 we present the network utility maximization problem with its constraints and utility function. Afterwards, in Section 3 we give the description of the various strategies for intentional node and edge attacks. Simulation results and analysis are given in Section 5 and Section 6 concludes this paper.

2 Network Utility Maximization Problem- NUM

Consider a network with m edges, labeled 1, . . . ,m, and n flows, labeled 1, . . . , n. Each flow has an associated nonnegative flow rate f_j; each edge or link has an associated positive capacity c_i. Each flow passes over a fixed set of links (its route); the total traffic t_i on link i is the sum of the flow rates over all flows that pass through link i. The flow routes are described by a routing matrix $R \in R^{mxn}$, defined as:

$$R_{ij} = \begin{cases} 1 & \text{flow } j \text{ passes through link } i \\ 0 & \text{otherwise.} \end{cases} \tag{1}$$

Thus, the vector of link traffic, $t \in R^m$, is given by $t = Rf$. The link capacity constraint can be expressed as $Rf \leq c$.

The aim of transmitting a flow of packets from their source to the destination is to get some benefit from the information transmission. Thus, it is natural to set a utility function U_j for flow j, and assume that U_i is related to its rate f_j. In this work as a utility function we use a function which provides proportional fairness among the end users:

$$U(f_j) = \log f_j \tag{2}$$

This function is strictly concave, because the second derivative is negative. From the concavity of the utility function it follows that the optimal rates $\{\hat{f}_j\}$ satisfy the following condition:

$$\sum_j \frac{f_j - \hat{f}_j}{\hat{f}_j} \leq 0, \tag{3}$$

This means that if rate of one transmitter rises, the rate of another transmitter will drop, and the drop will be proportionally larger than the rise. This property is known as the law of diminishing returns.

In order to maximize the utility we have to solve the following convex problem:

$$\begin{aligned} &\text{maximize } \textstyle\sum_{j=1}^{n} \log f_j \\ &\text{subject to } Rf \leq c, \end{aligned} \tag{4}$$

with variable f, and the implicit constraint $f \geq 0$.

Some comments about the NUM problem are given in the text below.

An unfair resource allocation is also possible, in which the goal is to maximize the overall throughput without any consideration about the fairness among the end users. If this is the case, then the unfair utility function would be:

$$U(f_j) = f_j \tag{5}$$

Additionally some reformulations and relaxations can be used by which the NUM problem can be decomposed both horizontally and vertically, and can be solved in distributed manner as in [16] and [17]. These decompositions are not needed for our analysis, because we are interested in overall network performance, so we solve the problem in a centralized manner.

In order to represent the performance of the complex network we use the maximum end-to-end throughput (*MT*) as performance indicator. *MT* is the total amount of bits received by all nodes per second and is measured in Mega bits per second (Mbps):

$$MT = \sum_{j \in n} f_j \tag{6}$$

3 Attack Strategies

In order to assess the vulnerability of the network we are considering two kind of intentional attacks: node and edge attack. In the network of computers attacks on nodes can be interpreted as breakdowns of servers by malicious hackers, while the attacks of edges may correspond to the cutting off the communication links. Additionally, the attacker can choose different strategies for node or edge removal, which are based on various centrality measures. These centrality measures can be based on the initial information about the network or on the information obtained by recalculation, when some of the nodes or edges are removed. We call the first ones *initial* and second ones *recalculated*. In the part below we will describe the different centrality measures that we are using for node or edge removal.

3.1 Degree Centrality – DEG

This measure is based on the idea that more important nodes (edges) are more active, that is, they have more neighbors in the graph [18,19]. It may be used for finding the core nodes (or edges) of a certain community. In order to use this measure for edge attack we are defining the edge degree k_e from the local information of the node degrees [14]:

$$k_e \equiv k_v k_w \tag{7}$$

where the edge e connects two nodes v and w with node degrees k_v and k_w, respectively.

3.2 Betweenness Centrality – BTWN

This is a measure of the importance of a node in a network, and is calculated as the fraction of shortest paths between node pairs that pass through the node. Betweenness is, in some sense, a measure of the influence a node has over the flow of information through the network. Let G be a graph given with set of nodes V and set of edges E. Let s and t are be nodes of the graph. σ_{st} is the number of paths that pass from s to t. Let $\sigma_{st}(v)$ be the number of shortest paths that pass through the node v. The central betweenness of node v is:

$$C(v) = \sum_{s \neq v \neq t \in V} \frac{\sigma_{st}(v)}{\sigma_{st}} \tag{8}$$

Just like node betweenness denotes the importance of the nodes, the edge betweenness, in the similar way assigns values to links according to their importance. It is calculated as a number of shortest paths that pass through the edge. Let $\delta_{st}(e)$ be the number of shortest paths from s to t that pass through the edge e and δ_{st} be the total number of paths from s to t. The edge betweenness of edge e is:

$$C(e) = \sum_{s \neq t \in V} \frac{\delta_{st}(e)}{\delta_{st}} \tag{9}$$

3.3 Eigenvector Centrality (Pagerank) - PR

With this measure we can find out the importance of nodes according to the adjacent matrix of a connected graph [20,21]. It assigns relative scores to all nodes in the network based on the principle that connections to high-scored nodes contribute more to the score of a node than connections to low-scored nodes. In order to use this centrality measure for finding the importance of edges we first transform the node adjacency matrix into edge adjacency matrix and then we use the pagerank algorithm. The transformation is done in a way that we say that two links are neighboring if they are connected to the same node.

4 Simulation and Results

For our simulations we are using the above mentioned network models, where each network generator generates 5 samples of the 4 network models. Each sample has 100 nodes and average node degree around 6. The number of flows in each scenario is 1000 and each O-D (Origin – Destination) pair is generated randomly. The flow rate f_j is also generated randomly and it is between 0 and 1. The capacity c_i of the all links is equal to 1.

In order to solve our network utility maximization problem defined with (4) we are using CVX [25]. CVX is a modeling system for disciplined convex programming (DCP). DCP is a methodology for constructing convex optimization problems and is meant to support the formulation and construction of optimization problems that the user intends from the outset to be convex. DCP imposes a set of conventions or rules. Problems which follow the rules can be rapidly and automatically verified as convex and converted to solvable form. Some problems can be reformulated to be made convex and then solved by appropriate methods for convex problems.

The simulation starts with calculating the maximum end-to-end throughput (6) for the given network. Afterwards, we attack a certain node (or edge) by removing it from the network, using one of the mentioned strategies in Section 3. The flows which originate or end at this node are randomly transferred to a different node, while the flows which go through the node reconfigure their routes to find new shortest paths. The removal of the nodes (or links) changes the entries in the adjacency matrix. Using the new routing information we compute a new optimal bandwidth allocation using (4). In these simulation scenarios we use static routing, that means that we do not take into consideration the load balancing. In addition, the simulation for a given network stops when the network falls apart into two or more islands.

In the next part we will show and analyze some of the interesting results we have obtained in our simulations.

In Fig. 1 we show the flow for the ER when nodes are attacked with the suggested recalculated strategies. One can see that the flow is decreasing in the same fashion for all the three strategies. The only difference between the strategies is that the pagerank disconnects the network when the smallest number of nodes is removed. By removing 8% of the nodes the flow decreased by 43%.

Fig. 2 shows the same analysis only now strategies based on initial information are used. We can see quite interesting phenomenon, i.e. by removing the third most important node, the flow increases instead of decreasing. For the explanation of this phenomenon refer to [14]. Additionally, with these attacks based on initial information when removing 8% of the nodes the flow decreased by around 37%.

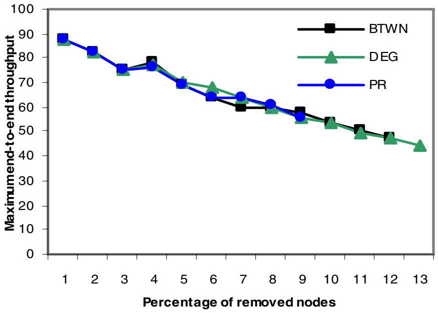


Fig. 1. Maximum end-to-end throughput for the random networks (ER) when attacking nodes using recalculated strategies, such as: betweenness centrality (BTWN), degree centrality (DEG) and pagerank (PR)

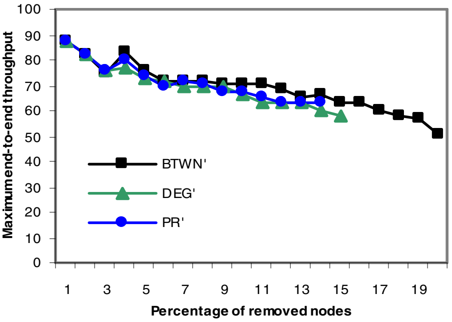


Fig. 2. Maximum end-to-end throughput for the random (ER) networks when attacking nodes using strategies based on initial information, such as: betweenness centrality (BTWN), degree centrality (DEG) and pagerank (PR)

For the scale-free networks the recalculated strategies for attack gave the same performance and they disconnect the network only when 5% of the most important nodes were removed. For the attacks based on initial information, from comparing fig. 3 with figures 1, 2, 4, and 5, one can see that the decreasing slop of the flow curve is much bigger than for the rest of the networks.

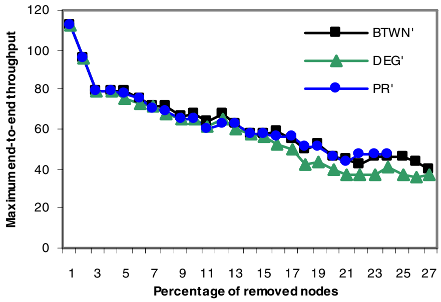


Fig. 3. Maximum end-to-end throughput for the scale-free (SF) networks when attacking nodes using recalculated strategies, such as: betweenness centrality (BTWN), degree centrality (DEG) and pagerank (PR).

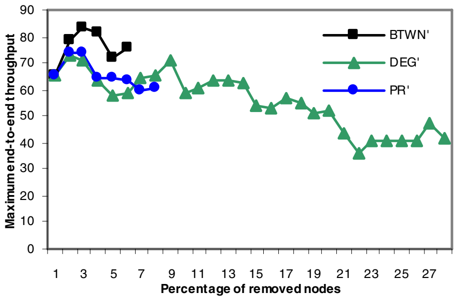


Fig. 4. Maximum end-to-end throughput for the random geometric (GR) networks when attacking nodes using strategies based on initial information, such as: betweenness centrality (BTWN), degree centrality (DEG) and pagerank (PR)

Fig. 4 shows how the strategies based on initial information affect the flow in GR networks. It is noticeable that BTWN disconnects the network when the smallest number of nodes is removed. After which came pagerank and the degree strategy needs around 27% of the nodes in order to disconnect the network. In addition, the slope of the flow curve is the smallest, which means that this kind of attacks does not reduce the flow too much, like in the other networks. From Fig. 4 we can notice the same phenomenon, mentioned before, i.e. by removing certain nodes the maximum-end-to-end throughput increases instead of decreasing.

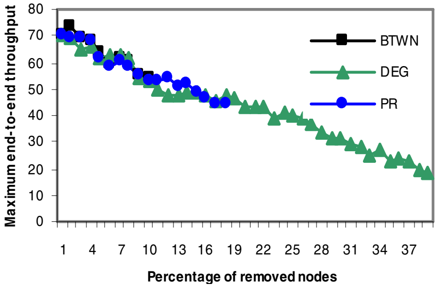


Fig. 5. Maximum end-to-end throughput for the small world (SW) networks when attacking nodes using recalculated strategies, such as: betweenness centrality (BTWN), degree centrality (DEG) and pagerank (PR).

For the small world network when using recalculated strategies by removing 10% of the nodes, the flow decreased only 27%, while when using strategies based on initial information it only decreased for 12% (Fig. 5). This means that this type of network is resistant to intentional node attacks when it comes to measuring the flow in the network. The three types of attack influenced the flow in the same manner. The only difference is that network was disconnected firstly with BTWN, then PR and lastly with DEG.

In Fig. 6 we show the total flow of the four types of network when using PR, based on initial information, as strategy for intentional node attack. The total flow in the networks before removing any node depends on the type of the network. For instance, the highest flow has the SF model, then the ER, SW and the last is the GR model. These results are equal with the results obtained when using game theory and the Method of Successive Averages as a technique for calculating the network vulnerability [26].

The total flow changes dramatically when we start to attack nodes based on the PR technique based on initial information. The highest performance drop has the SF and ER networks. The problem with the GR networks is that they can be easily broken into several disconnected regions (by removing about 1% of the total number of nodes).

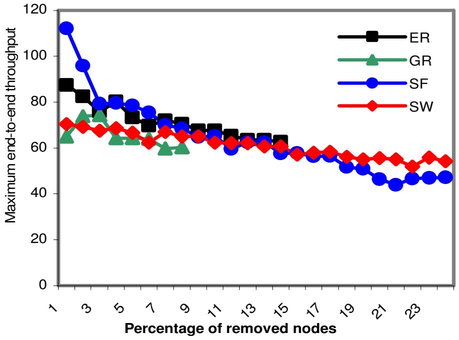


Fig. 6. Maximum end-to-end throughput for all synthetic complex networks when attacking nodes using pagerank based on initial information (PR)

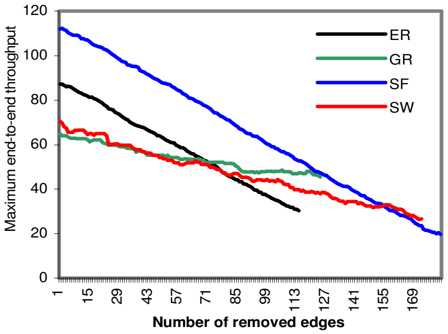


Fig. 7. Maximum end-to-end throughput for all synthetic complex networks when attacking edges using recalculated pagerank (PR)

We encountered almost the same results when instead of nodes we were attacking edges using the recalculated pagerank algorithm (see Fig. 7). It is noticeable that SF and ER networks at the beginning show the best performance, but after removing some of the edges (around 12%) the SW and GR networks outperform the ER network. Then when we continue to remove more edges (around 21%) the GR network performance is close to that of the SF network. At the end when we removed around 27% of the edges the SW outperforms the rest of the networks, when we measure the maximum end-to-end-throughput. When we removed 15% of the edges, the highest drop in the flow performance showed the ER network (around 43%), then SF (around 42%), then SW (around 36%) ant the lowest drop GR (around 26%). In order to disconnect the network, by attacking the edges with the PR strategy, the most robust to attacks was the SF (around 30% of the edges were needed to disconnect the network), SW (around 28%), ER (around 22%) and GR (around 18%).

5 Conclusion

This brief has studied the attack (node and edge) vulnerability of the different models for complex networks when the maximum end-to-end throughput of the network was taken into consideration. All of the models for complex networks show a considerable decline in performance when they encounter an intentional node or edge attack. One of the obtained results show that the scale-free networks have the highest maximum end-to-end throughput, but when removing nodes or edges the throughput decreases dramatically. The sharp decrease was also the case in the random networks. Additionally, it was shown that among the suggested recalculated and strategies based on initial information there is no big difference when we measure the throughput, whereas they differ in the percentage of nodes (or edges) needed to be removed in order to disconnect the network. The throughput is decreased more when instead of strategies based on initial information we are using recalculated strategies.

As a future work instead of static routing we want to use dynamic routing with load balancing, which takes into account the current flow in the edges, and by that we want to obtain more realistic results. Another improvement would be, instead of removing nodes (or edges), to use certain nodes to generate jam traffic in the network in order to reduce the maximum end-to-end throughput in the network, which presents a more realistic scenario than to remove some important node (or edge) in the network, which can be highly secured and protected.

References

1. Erdős, P., Rényi, A.: On the evolution of random graphs. Publ. Math. Inst. Hung. Acad. Sci. 5, 17–61 (1960)
2. Barabási, A.-L., Albert, R.: Emergence of scaling in random networks. Science 286, 509–512 (1999)
3. Penrose, M.: Random Geometric Graphs. Oxford University Press, New York (2004)
4. Newman, M.E.J.: The structure and function of complex networks. SIAM Rev. 45, 167–256 (2003)
5. Dezso, Z., Barabási, A.-L.: Halting viruses in scale-free networks. Phys. Rev. E 65, 055103 (2002)
6. Latora, V., Marchiori, M.: How the science of complex networks can help developing strategies against terrorism. Chaos, Solitons and Fractals 20, 69–75 (2004)
7. Nekovee, M., et al.: Theory of rumour spreading in complex social networks. Physica A 374, 457–470 (2007)
8. Kempe, D., Kleinberg, J.M., Tardos, E.: Maximizing the spread of influence through a social network. In: ACM SIGKDD international conference on Knowledge discovery and data mining (2003)
9. Checco, P., Biey, M., Vattay, G., Kocarev, L.: Complex network topologies and synchronization. In: Proc. ISCAS 2006, Kos, Greece, May 2006, pp. 2641–2644 (2006)
10. Crucitti, P., Latora, V., Marchiori, M.: Model for cascading failures in complex networks. Phys. Rev. E 69, 045104(R) (2004)
11. Guimera, R., Diaz-Guilera, A., Vega-Radondo, F., Cabrales, A., Arenas, A.: Optimal network topologies for local search with congestion. Phys. Rev. Lett. 89, 248701–248704 (2002)

12. Boccaletti, S., Latora, V., Moreno, Y., Chavez, M., Hwang, D.-U.: Complex networks: Structure and dynamics. Phys. Rep. 424, 175–308 (2006)
13. Motter, A.E., Lai, Y.-C.: Cascade-based attacks on complex networks. Phys. Rev. E 66, 065102(R) (2002)
14. Holme, P., Kim, B.J., Yoon, C.N., Han, S.K.: Attack vulnerability of complex networks. Phys. Rev. E 65, 056109 (2002)
15. Mirchev, M., Filiposka, S., Trajkovski, N., Trajanov, D.: Network utility maximization in ad hoc networks with different communication patterns. In: ETAI 2009, Ohrid, Macedonia (2009)
16. Kelly, F.P., Maulloo, A.K., Tan, D.K.H.: Rate control in communication networks: shadow prices, proportional fairness and stability. J. Optical Research Society 49, 237–252 (1998)
17. Kunniyur, S., Srikant, R.: End-to-end congestion control schemes: Utility functions, random losses and ECN marks. IEEE/ACM Transactions on networking 11(5), 689–702 (2003)
18. Freeman, L.: Centrality in social networks: Conceptual clarification. Social Networks 1(3), 215–239 (1979)
19. Nieminen, J.: On the centrality in a graph. Scandinavian Journal of Psychology 15(1), 332–336 (1974)
20. Bonacich, P.: Factoring and weighting approaches to status scores and clique identification. Journal of Mathematical Sociology 2(1), 113–120 (1972)
21. Larry, P., Sergey, B., Motwani, R., et al.: The PageRank citation ranking: Bringing order to the web (1998), http://citeseer.nj.nec.com/page98pagerank.html [04.06. 2003]
22. Karonski, M., Rucinski, A.: The Origins of the Theory of Random Graphs. The Mathematics of Paul Erdos. Springer, Berlin (1997)
23. Amaral, L.A.N., Scala, A., Barthelemy, M., Stanley, H.E.: Classes of small-world networks. PNAS 97, 11149–11152
24. Barabasi, A.L.: Linked. Penguin Group, London (May 2003)
25. CVX: Matlab Software for Disciplined Convex Programming, http://www.standofd.edu/~boyd/cvx
26. Igor, M., Filiposka, S., Gramatikov, S., Trajanov, D., Kocarev, L.: Game Theoretic Approach for Discovering Vulnerable Links in Complex Networks. In: International Joint Conferences on Computer, Information, and System Sciences, and Engineering, University of Bridgeport, USA, December 5-13 (2008)

Outage Probability of Multi-hop Relay Systems in Various Fading Channels

Jovan Stosic[1] and Zoran Hadzi-Velkov[2]

[1] Makedonski Telekom, Orce Nikolov bb, 1000 Skopje, Macedonia
jovan.stosic@telekom.mk
[2] Ss. Cyril and Methodius University, Faculty of Electrical Engineering
and Information Technologies, Karpos 2 bb, 1000 Skopje, Macedonia
zoranhv@feit.ukim.edu.mk

Abstract. In this paper we study the end-to-end outage performance of multi-hop cooperative communication systems employing amplify and forward (AaF) relaying under Rayleigh, Nakagami, Rician and Weibull fading channels. The outage probability performances of multi-hop systems with fixed gain and variable gain relays is compared. The outage probability for multi-hop systems under Rayleigh, Nakagami and Weibull fading models can be determined only by combining analytical results with numerical integration techniques. We show that fixed gain system has a better outage performance compared to the variable gain for all fading scenarios. This performance gap increases by increasing the number of hops.

Keywords: Wireless cooperative communications**,** outage probability, multipath fading, multi-hop relay systems.

1 Introduction

The newest trend in the contemporary wireless networks is based on the paradigm of partner cooperation, which has already occupied an entire new area of research in the wireless communications, called cooperative communications. Cooperative terminals exploit the properties of the multipath transmission of the radio signal in order to increase the efficiency and robustness of their communication. That means that the neighboring wireless stations, which are in the area of one transmitter-receiver pair, are “assisting” the communication between them in their “leisure time” by performing the function of relay, thus achieving the effect of virtual diversity (e.g. virtual multi-antenna array). The effect of spatial diversity is generated because the cooperating partners on different locations are resending independent copies of the signal over orthogonal channels. The receiver is combining the signal’s replicas from the source and other partners in a constructive manner in order to provide reliable decision of the transmitted symbol.

In such cooperative scenario, the outage probability (OP) is the key performance measure of the cooperative relaying system. In this paper, we study the end-to-end OP performance of the multi-hop relay systems operating over independent Rayleigh, Nakagami, Rician and Weibull fading channels. We consider amplify and forward (AaF) multi-hop systems with fixed and variable gains. A fixed gain relay requires knowledge of the average fading signal-to-noise ratio (SNR) of the previous hop, while the variable gain relay requires knowledge of the instantaneous channel state information (CSI) of each hop.

The remainder of this paper is organized as follows. Next Section presents the system model and presents our novel analytical results. Numerical results are presented in Section 3, and Section 4 concludes the article.

2 System and Channel Models

Fig. 1 presents the studied non-regenerative multi-hop communication system, which consists of the source *T*, the destination *D* and (*N*–1) AaF relays. Each hop is subjected to the independent but non-identical Rayleigh, Nakagami, Rician or Weibull fading, for which the per-hop SNR γ is distributed according to the probability distribution functions (PDFs) given by [5]:

$$p_{ray}(\gamma)=\frac{1}{\bar{\gamma}}\exp\left(-\frac{\gamma}{\bar{\gamma}}\right), \tag{1}$$

$$p_{nak}(\gamma)=\frac{m^{m}\gamma^{m-1}}{\bar{\gamma}^{m}\,\Gamma(m)}\exp\left(-\frac{m\gamma}{\bar{\gamma}}\right), \tag{2}$$

$$p_{ric}(\gamma)=\frac{(1+K)}{\bar{\gamma}}\exp\left(-\frac{\gamma}{\bar{\gamma}}-K\right)I_0\left(2\sqrt{\frac{K(1+K)\gamma}{\bar{\gamma}}}\right), \tag{3}$$

$$p_{wei}(\gamma)=\frac{c}{2}\left(\frac{\Gamma\left(1+\frac{2}{c}\right)}{\bar{\gamma}}\right)^{\frac{c}{2}}\gamma^{\frac{c}{2}-1}\exp\left(-\left(\frac{\gamma}{\bar{\gamma}}\Gamma\left(1+\frac{2}{c}\right)\right)^{\frac{c}{2}}\right), \tag{4}$$

respectively, where *m* in (2) is the Nakagami fading parameter, $\bar{\gamma}$ is the average per-hop SNR, *K* in (3) is the Rician factor, *c* in (4) is the Weibull parameter, $I_0(\cdot)$ in (3) is the modified zero-th order Bessel function of first kind, and $\Gamma(\cdot)$ in (2) and (4) is the Gamma function. In case of AaF, relays amplify and then forward the received signal from the previous node, while, in case of decode and forward (DaF), the relays fully decode the received signal and then forward it to the next hop. The variable gain and fixed gain relaying is modeled according to concepts presented in [1] and [2], respectively.

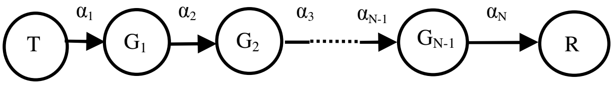


Fig. 1. *N*-hop system model

Fig. 1 presents the *N*-hop relaying scenario, with the (lowpass-equivalent) signal at the input of the *n*-th relay expressed as

$$r_n(t) = G_{n-1} \cdot \alpha_n \cdot \sqrt{\varepsilon_n} \cdot r_{n-1}(t) + w_n(t)\,, \tag{5}$$

where G_{n-1} is the gain of the (n-1)-th relay, α_n is the fading amplitude of the n-th hop, ε_n is the transmitting power of the (n-1)-th relay, r_{n-1} is the signal at the input of the (n-1)-th relay, and $w_n(t)$ is an additive white Gaussian noise (AWGN) with average power N_{0n}. Based on (5), the powers of the useful signal and the noise at the receiver output are respectively given by:

$$\begin{aligned} P_S &= (G_1^2 G_2^2 \cdots G_{N-1}^2)(\varepsilon_1 \alpha_1^2 \varepsilon_2 \alpha_2^2 \cdots \varepsilon_N \alpha_N^2)\,, \\ P_N &= (G_1^2 G_2^2 \cdots G_{N-1}^2)(\varepsilon_2 \alpha_2^2 \varepsilon_3 \alpha_3^2 \cdots \varepsilon_N \alpha_N^2) N_{01} + \\ &(G_2^2 \cdots G_{N-1}^2)(\varepsilon_3 \alpha_3^2 \varepsilon_4 \alpha_4^2 \cdots \varepsilon_N \alpha_N^2) N_{02} + \cdots + N_{0N}\,. \end{aligned} \tag{6}$$

By using (6), it can be shown that the signal-to-noise ratio (SNR) at the output of an *N*-hop multi-hop system, i.e., the receiver output, is expressed as

$$\gamma_{eq}^{-1} = \sum_{n=1}^{N} \left(\prod_{t=1}^{n-1} G_t^2 N_{0t} \prod_{t=1}^{n} \gamma_t \right)^{-1}. \tag{7}$$

For the case of variable gain relays, the *n*-th relay gain is set to [1]:

$$G_n^2 = \frac{1}{\alpha_n{}^2}\,, \tag{8}$$

where α_n is the *n*-th hop fading amplitude. In fixed gain case, the *n*-th relay gain is constant:

$$G_n = \frac{1}{C_n}\,, \tag{9}$$

where, according to [2], $C_n = \bar{\gamma}_n \left(e^{1/\bar{\gamma}_n} E_1(1/\bar{\gamma}_n) \right)^{-1}$. Note that $E_1(\cdot)$ is exponential integral function defined by [4, eq. (5.1.1)], and $\bar{\gamma}_n$ is average SNR of the *n*-th hop.

The OP is defined as the probability that the instantaneous output SNR falls below a predetermined threshold ratio γ_{th},

$$P_{out} = P[\gamma_{eq} < \gamma_{th}]. \tag{10}$$

To derive the analytical expression for the OP for the variable gain case, we used Moment Generating Function (MGF) based approach, devised in [1] and [5],

$$P_{out} = P(\gamma_{eq} < \gamma_{th}) = 1 - P(\frac{1}{\gamma_{eq}} < \frac{1}{\gamma_{th}}) = 1 - \mathcal{L}^{-1}(\frac{\mathcal{M}_{1/\gamma_{eq}}(-s)}{s})|_{1/\gamma_{th}} , \tag{11}$$

where $\mathcal{M}_{1/\gamma_{eq}}$ is MGF of the reciprocal of γ_{eq}. For Laplace transform inversion, we used Euler numerical technique given in [5, appendix 9B]:

$$P\left(\frac{1}{\gamma_{eq}} < \frac{1}{\gamma_{th}}\right) = \frac{2^{-k} e^{\frac{A}{2}}}{1/\gamma_{th}} \sum_{k=0}^{K} \binom{K}{k} \sum_{n=0}^{N+k} \frac{(-1)^n}{\alpha_n} \mathrm{Re}\left(\frac{M_{1/\gamma}\left(-\frac{A + 2\pi jn}{2/\gamma_{th}}\right)}{\frac{A + 2\pi jn}{2/\gamma_{th}}} \right) + E(A,K,N) , \tag{12}$$

$$E(A,K,N) = \frac{e^{-A}}{1 - e^{-A}} + \frac{2^{-k} e^{\frac{A}{2}}}{1/\gamma_{th}} \sum_{k=0}^{K} (-1)^{N+k+1} \binom{K}{k} \mathrm{Re}\left(\frac{M_{1/\gamma}\left(-\frac{A + 2\pi j(N+k+1)}{2/\gamma_{th}}\right)}{\frac{A + 2\pi j(N+k+1)}{2/\gamma_{th}}} \right) , \tag{13}$$

where $\alpha_n = 1$ for $n = 1,2,...,N$ and $\alpha_n = 2$ for $n = 0$. We set $A = 10\ln(10)$ in order to have discretization error less then 10^{-10}.

In case of Nakagami and Rayleigh (m_n=1) fading, we used the closed form expression for the reciprocal SNR of the n-th hop ($1/\gamma_n$):

$$\mathcal{M}_{1/\gamma_n}(-s) = \frac{2}{\Gamma(m)} \cdot \left(\frac{m_n s}{\bar{\gamma}_n}\right) K_{m_n}\left(2\sqrt{\frac{m_n s}{\bar{\gamma}_n}}\right) . \tag{14}$$

Under the assumption that the hops are subjected to independent fading and the variable relay gain is chosen according (8), the MGF of $1/\gamma_{eq}$ is product of the MGF's of $1/\gamma_n$, $n = 1,2,\ldots N$ [1].

In case of Rician PDF, it is impossible to find $\mathcal{M}_{1/\gamma_{eq}}$ in closed form, neither derive it with CAS (Computer Algebra System), therefore, we resorted to numerical integration. For the Weibull PDF, we modified PDF (4) in a more convenient form:

$$p_{\gamma wei}(\gamma) = bA^{-b}\gamma^{b-1} \exp\left(-\left(\frac{\gamma}{A}\right)^b\right) , \tag{15}$$

where A and b are given by

$$b = \frac{c}{2}; \quad A = \frac{\bar{\gamma}}{\Gamma\left(1 + \frac{2}{c}\right)} . \tag{16}$$

and then, we used symbolic integration for those values of the argument of s that are required by the Euler numeric technique, i.e.,

$$\mathcal{M}_{1/\gamma}(-s)=\int_{0}^{\infty} b\cdot A^{-b}\cdot\gamma^{b-1}e^{-\frac{s}{\gamma}-\left(\frac{\gamma}{A}\right)^{b}}d\gamma\ . \tag{17}$$

3 Numerical Results

In this Section, we present some illustrative figures which depict the excellent match between the numerical results and the corresponding results obtained by Monte Carlo simulation implemented in Matlab. For the dual-hop ($N = 2$) fixed gain relaying under Rayleigh fading, we use the closed form expression for the OP

$$P_{out}=1-2\sqrt{\frac{C\gamma_{th}}{\overline{\gamma}_1\overline{\gamma}_2}}\exp\left(-\frac{\gamma_{th}}{\overline{\gamma}_1}\right)K_1\left(2\sqrt{\frac{C\gamma_{th}}{\overline{\gamma}_1\overline{\gamma}_2}}\right). \tag{18}$$

where $K_1(\cdot)$ is first-order modified Bessel function of the second kind.

Fig. 2 presents the OP versus average per-hop SNR for dual-hop fixed gain system in Rayleigh fading. Note that $\overline{\gamma}_1=\overline{\gamma}_2$, and the thresholds γ_{th} are set as 0, 5 and 10dB respectively.

Fig. 3 presents the comparative curves of the OP of AaF system with fixed gain, AaF system with variable gain and a DaF system (used for comparison) with multiple

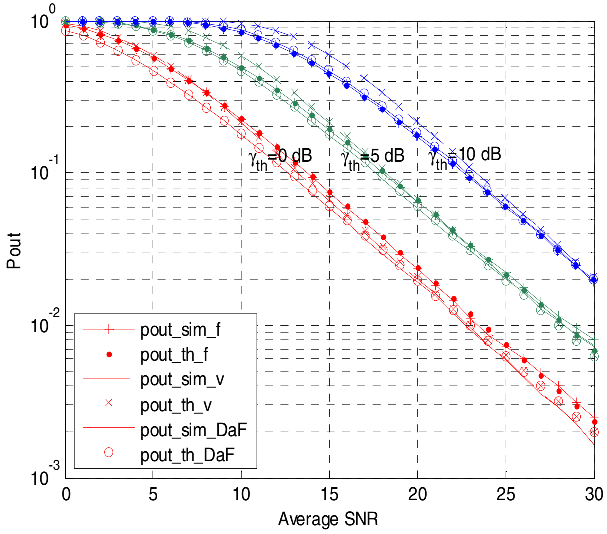


Fig. 2. OP performance of dual-hop system in Rayleigh fading with $\gamma_{th}=0, 5, \text{and } 10\ dB$

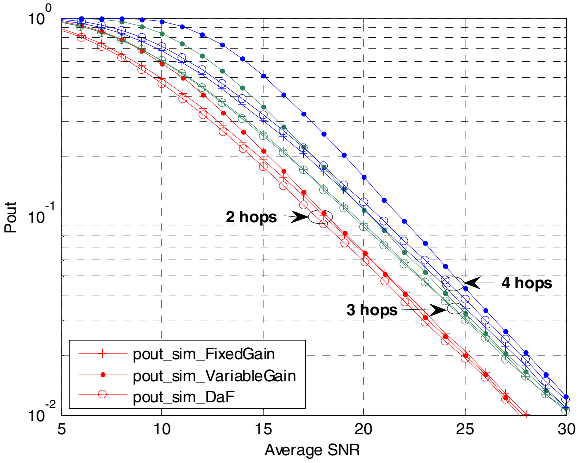


Fig. 3. OP for multi-hop system for N=2, 3 and 4 hops in Rayleigh fading when γ_{th} =5 dB

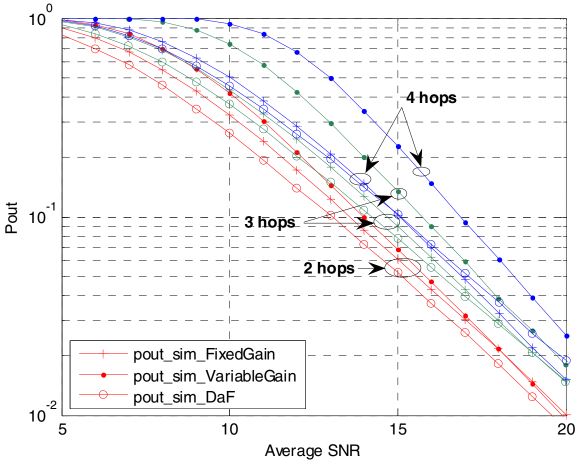


Fig. 4. OP for relaying system for N=2,3 and 4 in Weibull fading when γ_{th}=5 dB

hops in Rayleigh fading. Note that N = 2, 3 and 4, and $\gamma_{th} = 5\,dB$. It is obvious that, for the medium to large average SNRs, the two-hop systems with variable gain relays slightly outperform two-hop system with fixed gain relays, and for low to medium SNRs, the fixed gain systems outperform variable-gain systems. However, as the number of hops increase, the fixed-gain multi-hop systems outperform the corresponding variable gain multi-hop systems for all regions of SNR given identical fading model is applied. We obtained the similar results for other types of channels as well. The illustrative results for the case of the Weibull PDF are presented on Fig. 4. It is important to mention that in our analysis saturation effect of the fixed gain relays is not taken in to account.

We also analyzed the multi-hop system with 4 hops and thresholds SNRs γ_{th} of 0, 5 and 10 dB under various types of fading. In Fig. 5 we present the results for a four-hop system in case when all hops are subjected to Rayleigh fading, whereas in Fig. 6 we present results for Nakagami fading. Fig. 7 and Fig. 8 present the results for Rician and Weibull fading, respectively. For this numerical analysis, we set the PDF parameters to following values: *m*=2.285, *K*=3 and *b*=1.5. In all cases for $\gamma_{th} = 0$ (except for Rician fading), results again indicate that, for medium to large average SNRs, the variable gain systems slightly outperform the fixed gain systems. However,

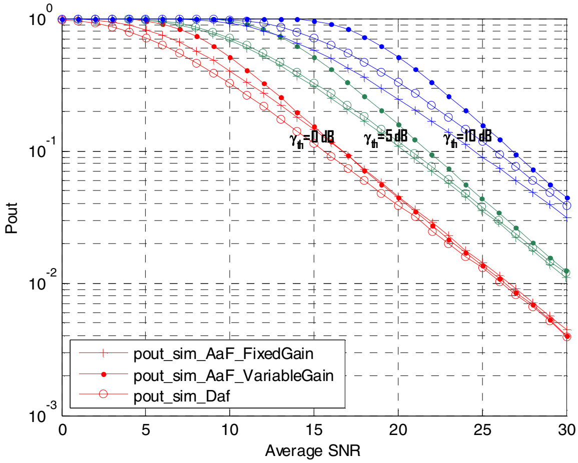


Fig. 5. OP for 4-hop system in Rayleigh fading γ_{th} = 0, 5, and 10 *dB*

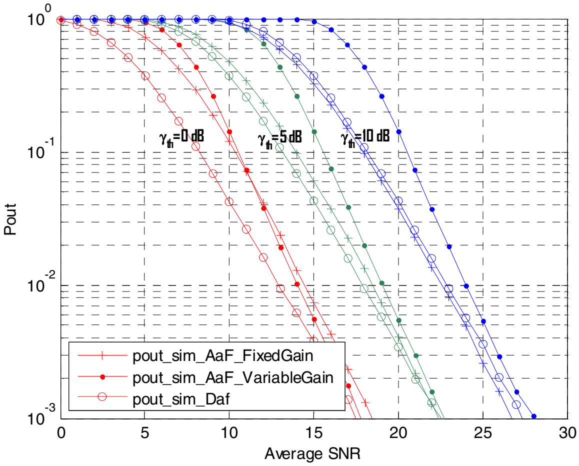


Fig. 6. OP for 4 hop system in Nakagami fading where γ_{th}=0, 5, 10 dB, and m= 2.3

for $\gamma_{th} > 0$, we notice better performance of the fixed gain systems for all channel models. For higher γ_{th}, the performance gap between the two systems is increased in favor of the fixed gain systems.

For 4-hop systems in Rician fading, the fixed gain systems outperform the variable gain system in all SNR regions regardless of the thresholds γ_{th} (Fig. 7). The gap in OP performance increases as γ_{th} increases. We emphasize that the fixed gain relays achieve their best possible performance since the effect of the saturation sensibility of fixed gain system is not accounted for. Moreover, in all cases it can be observed that as the number of hops increase fixed gain system even slightly outperform DaF system due to sub-optimal power allocation in DaF systems [3].

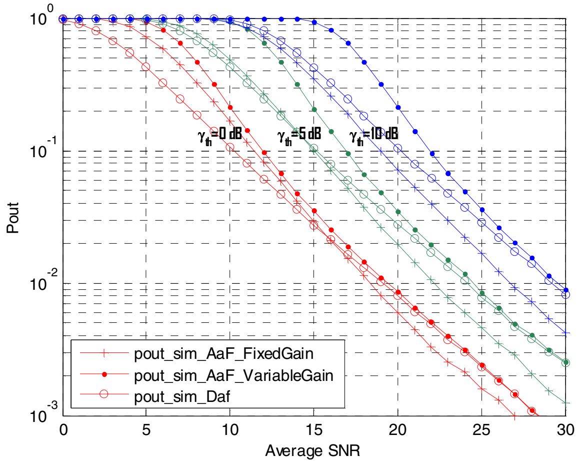


Fig. 7. OP for 4 hop system in Rician fading where γ_{th}=0, 5, 10 dB, and K=3

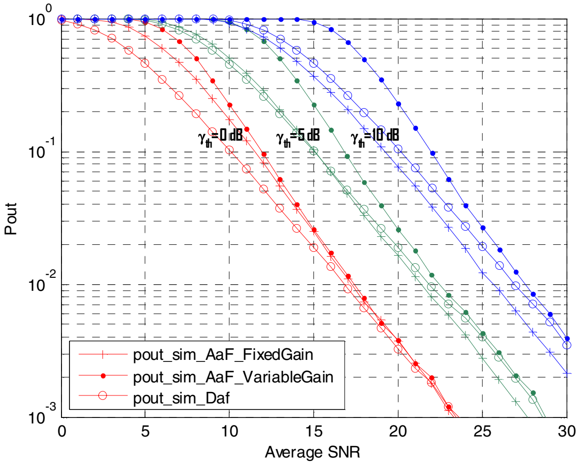


Fig. 8. OP for 4 hop system in Weibull fading (γ_{th}=0, 5, 10 dB))

In Fig. 9 we compared the OP performances of 4-hop systems under Nakagami fading with fading parameter *m*= 2.2857 and Rician fading with the respective Rician factor *K*=3, calculated according expression (19). With such selection of the parameters similar fading effects should be expected.

$$K = \frac{\sqrt{m^2 - m}}{m - \sqrt{m^2 - m}} . \tag{19}$$

Considering Fig. 9 we find out that such approximation is relatively successful since the OP performances for the two channel models are similar, particularly for low to medium average SNRs. As average SNR increases, the Nakagami fading model foresees lower OP compared to the Rician fading model.

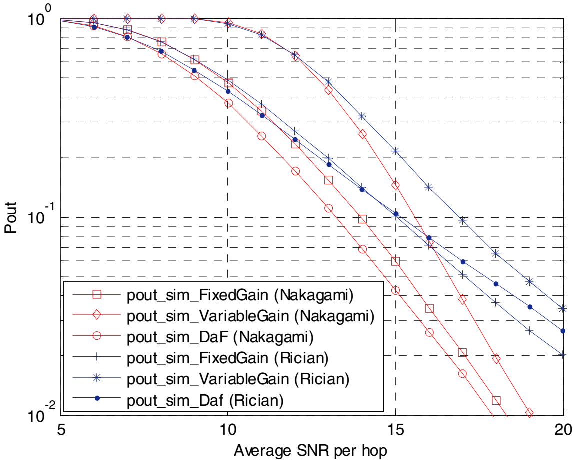


Fig. 9. OP for 4 hop system in Nakagami and Rician fading (γ_{th}= 5 dB)

4 Conclusion

In this paper, the outage performance of multi-hop AaF wireless relay system with fixed and variable gain relays had been studied by combination of analytical, numerical and simulation methods. Due to the complexity of the output SNR, the OP for multi-hop systems under Rayleigh, Nakagami, Rician and Weibull fading models can be determined only by combining analytical results with numerical integration techniques, except for the dual hop case.

Despite their lower complexity, the multi-hop system with fixed gain relays typically outperforms the variable gain system. The performance gap between these two systems increases substantially by increasing the number of hops for all considered fading channels regardless of the selection of the average per hop SNR. The increase of the SNR threshold further widens this gap.

References

1. Hasna, M.O., Alouini, M.S.: Outage Probability of Multihop Transmission Over Nakagami Fading Channels. IEEE Communications Letters 7(5) (May 2003)
2. Hasna, M.O., Alouini, M.S.: A Performance Study of Dual-Hop Transmissions With Fixed Gain Relays. IEEE Transactions on Wireless Communications 3(6) (November 2004)
3. Hasna, M.O., Alouini, M.S.: Optimal Power Allocation for Relayed Transmissions Over Rayleigh-Fading Channels. IEEE Transactions on Wireless Communications 3(6) (November 2004)
4. Abramowitz, M., Stegun, I.A.: Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables, 9th edn. Dover, New York (1970)
5. Simon, M.K., Alouini, M.S.: Digital Communication over Fading Channels, 2nd edn. Wiley, New York (2005)
6. Sendonaris, A., Erkip, E., Aazhang, B.: User Cooperation Diversity Part I and Part II. IEEE Trans. Commun. 51(11), 1927–1948 (2003)
7. Nosratinia, A., Hunter, T.E., Hedayat, A.: Cooperative Communication in Wireless Networks. IEEE Communications Magazine (October 2004)
8. Liu, P., Tao, Z., Lin, Z., Erkip, E., Panwar, S.: Cooperative wireless communications: A cross-layer approach. IEEE Wireless Communications (August 2006)
9. Laneman, J.N., Wornell, G.W.: Exploiting Distributed Spatial Diversity in Wireless Networks. In: Proc. 40th Allerton Conf. Communication, Control, Computing, Allerton Park, IL, September 2000, pp. 775–785 (2000)
10. Prabhu, G.S., Shankar, P.M.: Simulation of flat fading using MATLAB for classroom instruction. IEEE Transactions on Education 45(1), 19–25 (2002)
11. Oyman, Ö., Laneman, J.N., Sandhu, S.: Multihop relaying for broadband wireless mesh networks: From theory to practice. IEEE Communications Magazine 45(11), 116–122 (2007)
12. Spencer, Q.H., Peel, C.B., Swindlehurst, A.L., Haardt, M.: An introduction to the multi-user MIMO downlink. IEEE Communications Magazine 42(10) (October 2004)

Non-poisson Processes of Email Virus Propagation

Miroslav Mirchev[1] and Ljupco Kocarev[1,2,3]

[1] Faculty of Electrical Engineering and Information Technologies, Skopje, Macedonia
[2] Macedonian Academy for Sciences and Arts, Skopje, Macedonia
[3] University of California, San Diego, CA
{miroslavm,lkocarev}@feit.ukim.edu.mk

Abstract. Email viruses are one of the main security problems in the Internet. In order to stop a computer virus outbreak, we need to understand email interactions between individuals. Most of the spreading models assume that users interact uniformly in time following a Poisson process, but recent measurements have shown that the intercontact time follows heavy-tailed distribution. The non-Poisson nature of contact dynamics results in prevalence decay times significantly larger than predicted by standard Poisson process based models. Email viruses spread over a logical network defined by email address books. The topology of this network plays important role in the spreading dynamics. Recent observations suggest that node degrees in email networks are heavy-tailed distributed and can be modeled as power law network. We propose an email virus propagation model that considers both heavy-tailed intercontact time distribution, and heavy-tailed topology of email networks.

Keywords: Computer viruses, Dynamical systems, Complex networks.

1 Introduction

The concept of a computer virus is relatively old in the young and expanding field of information security. It was first developed by Cohen in [1, 2], and it still is an active research area. Computer viruses still accounts for a significant share of the financial losses that large organizations suffer for computer security problems, and it is expected that future viruses will be even more hostile.

According to The WildList Organization International [3] there were 70 widespread computer viruses in July 1993, and that number have increased up to 953 in July 2009 (Fig. 1). With the proliferation of broadband ''always on'' connections, file downloads, instant messaging, Bluetooth-enabled mobile devices, and other communications technologies, the mechanisms used by viruses to spread have evolved as well [4, 5]. Still, many viruses continue to spread through email. Indeed, according to the Virus Bulletin [6], the email viruses (email worms) still accounts for large share of the virus prevalence today.

Email viruses spread via infected email messages. The virus may be in an email attachment or the email may contain a link to an infected website. In the first case the virus will be activated when the user clicks on the attachment and in the second case when the user clicks on the link leading to the infected site.

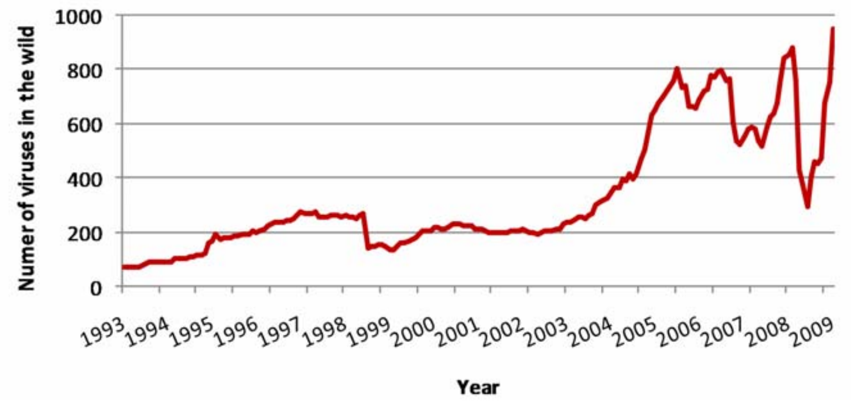


Fig. 1. Number of viruses in-the-wild according to The WildList Organization International. This is a cooperative listing of viruses reported as being in the wild by virus information professionals. The list includes viruses reported by multiple participants, which appear to be non-regional in nature. The WildList is currently being used as the basis for in-the-wild virus testing and certification of anti-virus products by the ICSA, Virus Bulletin and Secure Computing.

When an email virus infects a machine, it sends an infected email to all addresses in the computer's email address book. This self-broadcast mechanism allows for the virus's rapid reproduction and spread, explaining why email viruses continue to be one of the main security threats. While some email viruses used only email to propagate (e.g. Melissa), most email viruses can also use other mechanisms to propagate in order to increase their spreading speed (e.g. W32/Sircam, Love Letter).

Although virus spreading through email is an old technique, it is still effective and is widely used by current viruses. It is attractive to virus writers, because it doesn't require any security holes in computer operating systems or software, almost everyone uses email, many users have little knowledge of email viruses and trust most email they receive (especially email from friends) and email is private property so correspondent laws or policies are required to permit checking email content.

Email viruses usually spread by connecting to SMTP servers using a library coded into the virus or by using local email client services. Viruses collect email addresses from victim computers, in order to spread further, by: scanning the local address book, scanning files with appropriate extensions for email address and sending copies of itself to all mail in the user's mailbox. Some viruses even construct new email addresses with common domain names.

In order to eradicate viruses, as well as to control and limit the impact of an outbreak, we need to have a detailed and quantitative understanding of the spreading dynamics and environment. In most email virus models have been assumed that the contact process between individuals follows Poisson statistics, and, the time between two consecutive contacts is predicted to follow an exponential distribution [7-13]. Therefore, reports of new infections should decay exponentially with a decay time of about a day, or at most a few days [7–11]. In contrast, prevalence records indicate that new infections are still reported years after the release of antiviruses [4, 7, 14], and their decay time is in the vicinity of years, 2–3 orders of magnitude larger than the Poisson process predicted decay times. This discrepancy is rooted in the failure of the Poisson approximation for the interevent time distribution. Indeed, recent studies of email exchange records between have shown that the probability density function of the time interval between two consecutive emails sent by the same user is well approximated by a fat tailed distribution [15-19]. In [20] the authors prove that this deviation from the Poisson process has a strong impact on the email virus's spread, offering a coherent explanation of the anomalously long prevalence times observed for email viruses.

The email network is determined by users' email address books, and its topology plays important role in the spreading dynamics. In [21] the authors use Yahoo email groups to study the email network topology. Although the topology of email groups is not the complete email network topology, they use it to figure out what the topology might be like. Their findings suggest that the email groups are heavy-tailed distributed, so it is reasonable to believe that email network is also heavy-tailed distributed. The problem of virus spreading in networks with heavy-tailed distribution has been studied in [7, 10, 21].

An epidemic threshold is a critical state beyond which infections become endemic. In [22, 23], the authors have presented a model that predicts the epidemic threshold of a network with a single parameter, namely, the largest eigenvalue of the adjacency matrix of the network.

In this paper, we propose an email virus propagation model with nonlinear dynamical system, which considers both heavy-tailed intercontact time distribution and heavy-tailed topology of email networks. We use this model to reveal new form of the epidemic threshold condition.

The rest of the paper is organized as follows. In Section 2, we define the network model, and analyze the email network topology and communication patterns. After that in Section 3, we propose a discrete stochastic model for Non-Poisson virus propagation in email networks with power law topology and have-tail distributed interevent times. Simulation results and analyses are given in Section 4 and Section 5 concludes the paper.

2 Email Network Model

Let $G = (V, E)$ be a connected, undirected graph with N nodes, which represent the email users, and m edges, which represent the contacts between the users. Every user has an address book in which he has all the users he contacts with. These address books are represented with the adjacency matrix $\mathbf{A}$ of the graph G, i.e., $a_{ij} = 1$ if $(i, j) \in E$ (user i have user j in his address book) and $a_{ij} = 0$ otherwise.

At time k, each node i can be in one of two possible states: **S** (susceptible) or **I** (infected). The state of the node is indicated by a status vector which contains a single 1 in the position corresponding to the present status, and 0 in the other position:

$$\mathbf{s}_i(k) = [s_i^S(k) \;\; s_i^I(k)]^T \tag{1}$$

and let

$$\mathbf{p}_i(k) = [p_i^S(k) \;\; p_i^I(k)]^T \tag{2}$$

be the probability mass function of node *i* at time *k*. For every node *i* it states the probability of being in each of the possible states at time *k*.

The network topology is determined by the adjacency matrix **A**, i.e. by the users' email address books. The size of a user's email address book is the degree of the corresponding node in the network graph. Since email address books are private property, it is hard to find data to tell us what the exact email topology is like.

We use the Enron email dataset, described in [24] and available at [25], to study the email network topology. This set of email messages was made public during the legal investigation concerning the Enron Corporation. It is the only publicly available email dataset and consists of 158 users (mostly senior management) and 200,399 messages (from which 9728 are between employees). The dataset contains messages from a period of almost three years. On Fig. 2 the degree distribution of the users' address books from this dataset is shown. We see that the power law $P(k) \sim k^{-3.5}$ approximates well a substantial part of the users' degree distribution, but fails to approximate well for small degree values. This is mostly due to the fact that the number of users in the dataset is small, but nevertheless it gives us an insight into the real email network topology. Because of this degree distribution, and the findings from [21], it is best if we model the email network as a power law network.

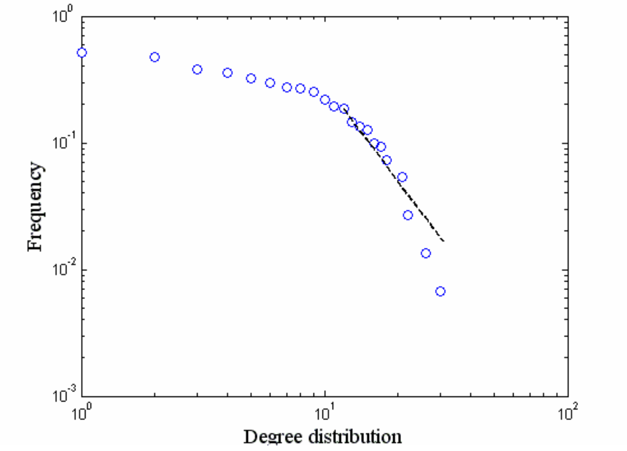


Fig. 2. Degree distribution of the address books from the Enron email dataset. The dashed line represents the power law decay $P(k) \sim k^{-3.5}$.

The contact dynamics responsible for the spread of email viruses is driven by the email communication and the usage patterns of individuals. To characterize these patterns we also use the Enron email dataset. We use only the messages between the employees (9728 messages), and it is sufficient for accurate analysis. Let τ (interevent time) denote the time between two consecutive emails sent by a single user. The distribution of the aggregate interevent of all the users approximately follows a power law with exponent $\alpha \approx 2.4$ and a cut-off at large τ values (Fig. 3).

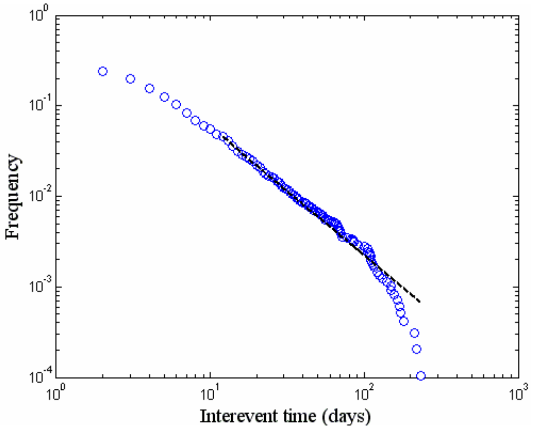


Fig. 3. Distribution of the interevent time between two consecutive emails sent by an email user in the Enron dataset. We aggregate the interevent times of all users (the distribution for single users is similar). The dashed line represents the power law decay $P(k) \sim k^{-2.4}$.

3 Email Virus Propagation Model

The spreading dynamics is jointly determined by the email activity patterns and the topology of the corresponding email communication network. We propose a discrete stochastic model for virus propagation in email network with power law topology and communication pattern with heavy-tailed interevent time distribution.

The Barabasi-Albert model [26] is used for generating email networks with power law topology, which is one of several proposed models that generate power law networks. The model is using a preferential attachment mechanism and generates network which has degree distribution with the power law form $P(k) \sim k^{-3}$.

In order to compare power law networks against random networks, we use the Erdos-Renyi model [27] for generating random networks. In this model, a graph $G(N, p)$ is constructed by connecting N nodes randomly. Each edge is included in the graph with probability p, with the presence or absence of any two distinct edges in the graph being independent.

When an email user have received message with a virus attachment by some of his contacts, he may discard the message (if he suspects the email or detects the email virus by using anti-virus software) or open the virus attachment if unaware of it. When the virus attachment is opened, the virus immediately infects the user and sends out virus email to all email addresses on this user's email address book. Different users open virus attachments with different probabilities, depending on their computer security knowledge. We assume that the probability that an email user opens the infected attachment, after he has received some infected message is constant and denote it with β. The infected user will not send out virus email again unless the user receives another copy of the email virus and opens the attachment again.

It takes time before a recipient receives a virus email sent out by an infected user, but the email transmission time is usually much smaller comparing to a user's email checking time. Thus in our model we ignore the email transmission time. In most cases received emails are responded to in the next email activity burst [15, 17], and viruses are acting when emails are read, approximately the same time when the next bunch of emails are written. According to this email users' activity can be represented as follows. Let $b_j(k)$ represent users' j activity at time k. If user j is active at time k $b_j(k) = 1$, otherwise $b_j(k) = 0$. We assume that a user reads all his emails at the moment he is active.

We model email users activity by using chaotic-maps. This method is used in [28, 29] for modeling packet traffic. The following map is convenient for our purposes:

$$x_j(k+1) = \begin{cases} \dfrac{x_j(k)}{(1-c_1 x_j(k)^{m_1-1})^{\frac{1}{m_1-1}}}, & \text{if } x_j(k) < d \\ 1 - \dfrac{1-x_j(k)}{(1-c_2(1-x_j(k))^{m_2-1})^{\frac{1}{m_2-1}}}, & \text{if } x_j(k) \geq d \end{cases} \tag{3}$$

where

$$c_1 = \frac{1-d^{m_1-1}}{d^{m_1-1}} \tag{4}$$

$$c_2 = \frac{1-(1-d)^{m_2-1}}{(1-d)^{m_2-1}}, \tag{5}$$

and $d \in [0, 1]$. At each time k, the value of $x_j(k)$ is evaluated for each user j, and then:

$$b_j(k) = \begin{cases} 0, & \text{if } x_j(k) < d \\ 1, & \text{if } x_j(k) \geq d \end{cases} \tag{6}$$

We choose this chaotic map, because for values of m_1 and/or m_2 in the range (3/2, 2) the map generates interevent times that have heavy tailed distribution. More precisely for d=0.7, m_1 = 1.53 and m_2 = 1.96 the distribution approximately follows a power law with exponent $\alpha \approx 2.4$ and a cut-off at large τ values, very similar to the true interevent time distribution (this can be achieved with other values also).

At the beginning (k = 0) there is a small number of initially infected users. Let $V(k)$ denote the infected inbox matrix, where $v_{ij}(k) = 1$, if user j have unread infected email message from user i at time k, and otherwise $v_{ij}(k) = 0$. At time $k = 0$, $v_{ij}(0) = 1$, if user j have initially infected user i in his address book, and otherwise $v_{ij}(0) = 0$. At each time k:

$$v_{ij}(k+1) = (a_{ij} h_i(k)) + v_{ij}(k)(1-h_i(k)))(1-b_j(k)) \tag{7}$$

$$h_{ij}(k+1) = \begin{cases} 1, & \text{if } s_i^I(k) = 0 \text{ and } s_i^I(k+1) = 1 \\ 0, & \text{otherwise} \end{cases} \tag{8}$$

Previously we assumed that a user reads all his emails at the moment he is active. So if user *j* is active at time *k* ($b_j(k) = 1$), all the messages from the infected inbox matrix *V* should be removed, $v_{ij}(k+1) = 0$ for all *i*.

We introduce another parameter *δ*, which represents the curing probability. After some user gets infected, he may use some means (such as virus removal tool) to remove the virus from his computer. As with *β* we assume constant curing probability among users. Having defined all this, the equations describing the evolution of our email virus propagation model are:

$$\begin{aligned} p_i^S(k+1) &= b_i s_i^S(k)(1-f_i(k)) + (1-b_i)s_i^S(k) + s_i^I\delta \\ p_i^I(k+1) &= b_i s_i^S(k) f_i(k) + s_i^I(k)(1-\delta) \\ s_i^T(k+1) &= Multirealize[\mathbf{p}_i^T(k+1)] \end{aligned} \tag{9}$$

where *Multirealize*[.] performs a random realization for the probability distribution given with $\mathbf{p}_i^T(k+1)$, and:

$$f_i(k) = 1 - \prod_{j=1}^{N}(1-\beta v_{ji}(k)) \tag{10}$$

4 Simulations and Analyses

For our simulations, we use email networks with 1000 nodes representing the email users and 3000 links representing the users' address books. First, we compare the spreading of email viruses in power law and random (Erdos-Renyi) network, by using both Poisson process approximation and true interevent distribution. For this simulation we use $\delta = 0$, because we are interested in the spreading dynamics, i.e. the number of new infections, instead of the total number of infected users. The other parameter values are d=0.7, $m_1 = 1.53$, $m_2 = 1.96$ and $\beta = 0.5$. From Fig. 4 we see that the spreading process in the power law email network evolves more rapidly than in random network, i.e. the number of new infections at the beginning is higher. If we compare the different interevent distributions, we see that the Poisson process approximation evolves much faster and the spreading process ends in one order of magnitude faster than in the true interevent time distribution. The number of new infections in power law networks, after the initial period, slightly deviates from exponential decay, while in random networks the decay is clearly exponential.

Predicting the epidemic threshold condition is an important part of a virus propagation model. In [22, 23] the authors predict the epidemic threshold with a single parameter $\lambda_{1,\mathbf{A}}$, the largest eigenvalue of the adjacency matrix **A** of the network. They prove that if an epidemic dies out, then it is necessarily true that:

$$\frac{\beta}{\delta} < \frac{1}{\lambda_{1,\mathbf{A}}}. \tag{11}$$

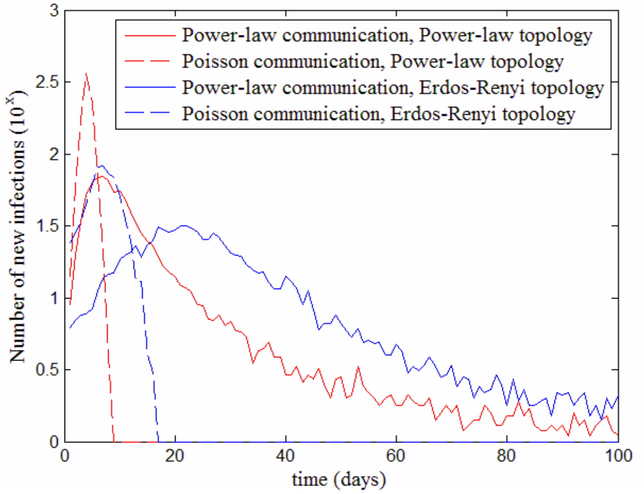


Fig. 4. Average number of new infections in Power-law network (red) and Erdos-Renyi (blue). After the initial period, the lines correspond to an exponential decay predicted by the Poisson process approximation (dash lines) and the true interevent distribution (solid lines).

The epidemic threshold in power law networks is zero [22], so we make the epidemic threshold analysis on random networks. We analyze the dependencies between the parameters, β, δ, $\lambda_{1,\mathbf{A}}$ and d at their threshold values, i.e. the values for which the system moves from a state where the virus prevails, to a state where the virus diminishes). The parameter d captures the characteristics of the communication pattern. We see (Fig. 5) that as in [22, 23] β and δ have linear dependency with $\lambda_{1,\mathbf{A}}$, while the threshold value of d exponentially increases, as $\lambda_{1,\mathbf{A}}$ increases. According to this, the epidemic threshold condition would have the form given in (12), which captures the essence of both network topology and communication patterns.

$$\frac{\beta}{\delta} < \frac{d^x}{\lambda_{1,\mathbf{A}}} \tag{12}$$

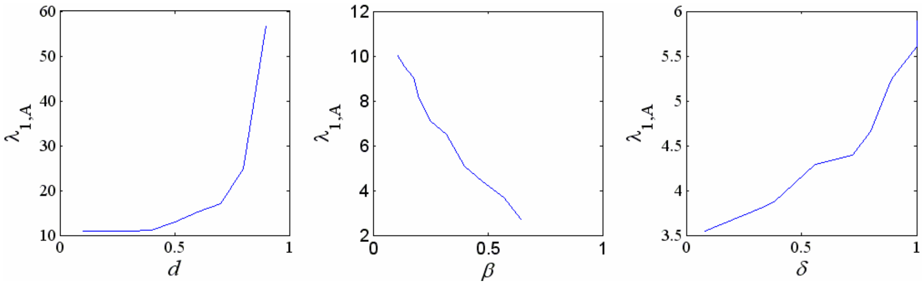


Fig. 5. The dependencies of the parameters β, δ, $\lambda_{1,\mathbf{A}}$ and d at the epidemic threshold.

5 Conclusion

In this paper we analyzed the email network topology and the email communication patterns. We proposed a model for virus propagation in email network with power law topology and communication pattern with heavy-tailed interevent time distribution. The analysis showed that the prevalence time for true interevent time distribution is much longer than predicted by standard Poisson based models, which is coincident with real data. Although the number of new infections exponentially decays in random networks, for email networks it slightly deviates from straight exponential decay.

The epidemic threshold analysis has revealed a new form of the condition under which an epidemic diminishes, which captures the essence of both network topology and communication patterns. This form will be further analyzed.

References

1. Cohen, F.: Computer Viruses. PhD Thesis, University of Southern California (1985)
2. Cohen, F.: Computer viruses – theory and experiments. Computers & Security 6(1), 22–35 (1987)
3. The WildList Organization International, `http://www.wildlist.org`
4. Wang, P., González, M.C., Hidalgo, C.A., Barabási, A.-L.: Understanding the Spreading Patterns of Mobile Phone Viruses. Science 324, 1071–1076 (2009)
5. Hu, H., Myers, S., Colizza, V., Vespignani, A.: WiFi networks and malware epidemiology. Proceedings of the National Academy of Sciences 106, 1318–1323 (2009)
6. Virus Bulletin, `http://www.virusbtn.com`
7. Pastor-Satorras, R., Vespignani, A.: Epidemic Spreading in Scale-Free Networks. Physical Review Letters 86(14), 3200–3203 (2001)
8. Meyers, L.A., Pourbohloul, B., Newman, M.E.J., Skowronski, D.M., Brunham, R.C.: Network theory and SARS: Predicting outbreak diversity. Journal of Theoretical Biology 232, 71–81 (2005)
9. Moreno, Y., Pastor-Satorras, R., Vespignani, A.: Epidemic outbreaks in complex heterogeneous networks. European Physical Journal 26, 521–529 (2002)
10. Barthélemy, M., Barrat, A., Pastor-Satorras, R., Vespignani, A.: Velocity and Hierarchical Spread of Epidemic Outbreaks in Scale-Free Networks. Physical Review Letters 92, 178701 (2004)
11. Boccaletti, S., Latora, V., Moreno, Y., Chavez, M., Hwang, D.U.: Complex Networks: Structure and Dynamics. Physics Report 424, 175–308 (2006)
12. Moreno, Y., Nevokee, M., Pacheco, A.F.: Dynamics of rumor spreading in complex networks. Physical Review E 69, 066130 (2004)
13. Nekovee, M., Moreno, Y., Bianconi, G., Marsili, M.: Theory of rumor spreading in complex social networks. Physica A 374(1), 457–470 (2007)
14. Pastor-Satorras, R., Vespignani, A.: Evolution and Structure of the Internet: A Statistical Physics Approach. Cambridge University, Cambridge (2004)
15. Eckmann, J.P., Moses, E., Sergi, D.: Entropy of dialogues creates coherent structures in e-mail traffic. Proc. of Natl. Acad. Sci. USA 101, 14333–14337 (2004)
16. Johansen, A.: Probing human response times. Physica A: Statistical Mechanics and its Applications 338, 286–291 (2004)

17. Barabasi, A.L.: Modeling bursts and heavy tails in human dynamics. Nature 435, 207–211 (2005)
18. Vazquez, A.: Impact of memory on human dynamics. Physica A: Statistical and Theoretical Physics 373, 747–752 (2007)
19. Vazquez, A.: Exact results for the Barabási Model of human dynamics. Physical Review Letters 95, 248701, 1–4 (2005)
20. Vazquez, A., Racz, B., Lukacs, A., Barabasi, A.L.: Impact of Non-Poissonian Activity Patterns on Spreading Processes. Physical Review Letters 98, 158702 (2007)
21. Zou, C., Towsley, D., Gong, W.: Email Virus Propagation Modeling and Analysis. Technical Report TR-CSE-03-04. Department of Electrical and Computer Engineering. Univ. of Massachusetts. Amherst
22. Wang, Y., Chakrabarti, D., Wang, C., Faloutsos, C.: Epidemic Spreading in Real Networks: An Eigenvalue Viewpoint. In: Proceedings of 22nd International Symposium on Reliable Distributed Systems, pp. 25–34 (2003)
23. Wang, Y., Chakrabarti, D., Wang, C., Leskovec, J., Faloutsos, C.: Epidemic Thresholds in Real Networks. ACM Transactions on Information and System Security (TISSEC) 10(4) (2008)
24. Klimt, B., Yang, Y.: Introducing the Enron Corpus. In: Proceedings of the 1st Conference on Email and Anti-Spam (CEAS 2004), Mountain View, CA (2004)
25. Enron Email Dataset, http://www.cs.cmu.edu/~enron/
26. Barabási, A.-L., Albert, R.: Emergence of scaling in random networks. Science 286, 509–512 (1999)
27. Erdős, P., Rényi, A.: The Evolution of Random Graphs. Magyar Tud. Akad. Mat. Kutató Int. Közl. 5, 17–61 (1960)
28. Erramilli, A., Roughan, M., Veitch, D., Willinger, W.: Self-Similar Traffic and Network Dynamics. Proceedings of the IEEE 90(5), 800–819 (2002)
29. Erramilli, A., Singh, R.P., Pruthi, P.: An application of deterministic chaotic maps to model packet traffic. Queueing Systems 20, 171–206 (1995)

Impact of Community Structures on Ad Hoc Networks Performances

Sonja Filiposka and Dimitar Trajanov

Faculty of Electrical Engineering and Information Technologies,
University of Ss. Cyril and Methodius, Karpos 2 bb., 1000 Skopje, R. Macedonia
{filipos,mite}@feit.ukim.edu.mk

Abstract. In this paper the impact of node communities on the ad hoc network performances is investigated. The community structures are viewed on the logical (application) level and on the physical level. They are modeled using complex network theory with which socially based mobility and communication patterns are developed. The different approaches in modeling offer the view of how the distribution of interconnections between the nodes influences the network performances. The results show that the logical interconnections have a great impact on the network performances especially in the cases when the node degree distribution follows the scale-free law.

Keywords: communities, ad hoc network, performances, complex networks, logical and physical level.

1 Introduction

The wireless fidelity of today's small and affordable devices have risen the massive need for instant and effortless networking on a whole new level that can not be satisfied with the usual access point configurations. What is expected from wireless devices today is the very definition of ad hoc networking: the ability to form a wireless mobile network anytime, anywhere without the use of any centralized administration or existing infrastructure [1].

Ad hoc networks offer this effortless networking by employing a special routing protocol that treats every device as end-node and a router at the same time. In this way the nodes act benevolently forwarding packets from other sources towards their destinations. This behavior allows for two nodes that are not in their radio range to communicate via their intermediate neighbor nodes forming so-called multihop paths.

However, the idea of creating an instant network as means for sharing information has influences on both, the networking protocols and the mobility and communication patterns. This means that as the need for the anytime, anywhere network establishing dictates the use of special ad hoc routing protocols, the same need of people to share information shapes the topology of the network as well as the traffic on the application level in the network.

The people that want to use the ad hoc network as a means for information sharing are part of an underlying social network that depicts the way people are interconnected.

The connections in the social network show the way people communicate by indicating the friendships and communities of friends with stronger ties inside the community. When we apply communication via ad hoc network over this social layout, it is normally expected that the network users will continue to honor the social ties and merely use a technological way to share the information between friends [2].

Thus, while we strive to tweak the ad hoc network performances so they can satisfy the need for higher throughput, it is essential that we model the network deployment using appropriate techniques. This means that while we investigate the network behavior, we must use communication and mobility models that will produce data and movement patterns that will reflect the social life of the network users.

In the light of this approach, we use the complex network theory in order to create a social model of the network users which will afterwards be used as an underlying basis for creating communication patterns and mobility scenarios in the network. When creating a model of a social network, there are few interesting characteristics that must be taken into account. It has been shown that all social networks which are part of the complex network family show certain global and local features that they have in common. This is the reason why complex network theory abandoned the social network modeling using the random graphs proposed by Erdos and Renyi, and moved on to the discoveries of the last decade which led to the small-world and scale-free properties.

We used the small-world and scale-free phenomena for modeling a socially based communication pattern for ad hoc networks and we investigated the impact this modeling had on the ad hoc network performances. The results presented in [3] and [4] show that the ad hoc network performances greatly differ when compared to the traditional way of modeling using a random communication pattern. This implies that the results obtained for the network performances behavior when investigating some special network aspects can be misleading. For example, the authors in [5] are investigating the performances of ad hoc networks that use different routing protocols on the basis of complete random traffic and random movement, which as we argued previously does not represent the real life deployment of the network. Similarly, in [6] the impact of coverage on the network performances in investigated using random models, but again this is under great influence of the social aspect of the network. In [7] performances of ad hoc networks using directional beams with group mobility are presented, and while the results include the physical aspects of social grouping, they also lack the application modeling of social information sharing.

However, one other characteristic of social networks is the natural community forming that occurs in the network. Social network exhibit the occurrence of a number of communities in the network that point the groups of friends that establish closer relationships inside their group while scarcely communicating outside the group via a small number of intergroup links. This forming of groups influences not only the communication pattern of having intense traffic inside the groups and lower amount of traffic between groups, but also the physical positioning of the nodes and their mobility pattern. The nodes that belong to the same group are expected to stay close to each other and the group is expected to share a common goal when moving.

The goal of this paper is to investigate how these community structures that occur in ad hoc networks because of their underlying social character influence the network performances. Of special interest is the question of how the link distribution density in the network, expressed via the social node degree distribution, affects the network performances in accordance with the community structures.

In order to pursue this topic we created several models of the logical and physical level of ad hoc nodes interactions creating socially based communication and mobility patterns that are socially based and reflect the small-world, scale-free and community forming phenomena of the network of users.

The rest of this paper is structured as follows: In section 2 the underlying social network models that incorporate the small-world phenomenon and the community structures, as well as the scale-free phenomenon and the community structures, are presented. Section 3 describes the way these network models are used for creating communication pattern and mobility scenarios for the ad hoc network performance investigation. In Section 4 the simulation scenarios are presented, while Section 5 gives an overview of the results obtained. Section 6 concludes the paper.

2 Modeling Social Networks of User Communities

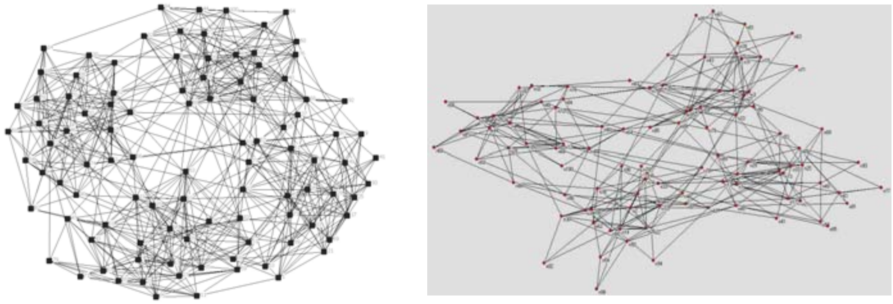
The complex network studies in the last decade have shown that social networks share some ubiquitous properties that have captured the interests of scientists from many different fields. These properties that have be found as an integral part of many natural and man-made networks are the small-world and scale-free phenomena.

The small world phenomenon [8] captures the so called low global separation in the network. Networks that exhibit the small-world phenomenon have a very small average path length when compared to the number of nodes in the network. For an example, the largest social network, i.e. mankind, has average path length of 5.8, the basis for the famous “six degrees of separation” statement.

The small-world phenomenon is successfully captured using the Watts and Strogatz model [9] that creates a network that has a large clustering coefficient (high local density), but also a small average path length. The large clustering coefficient is another feature that characterizes social networks implying the higher probability of two friends to have a mutual friend. However, one of the downfalls of the model is that the resulting network has a uniform distribution of the node degree, as well as the inability to create a network with a desired number of communities. In order to address the later problem we created a new small-world network model that allows for creation of a network with the same characteristics of the Watts-Strogatz model as well as the ability to form a given number of communities in the network. The small-world communities (SWcom) model is described in [3] and it emphasizes the existence of groups or clusters in the network wherein the nodes are tightly connected to each other while the network is holding by a number of inter-clusters links. That means that although the nodes tend to group, the inter-groups links create paths of just a few intermediate nodes to any destination in the network.

On the other hand, the scale-free phenomenon [10] is mainly concerned with the node degree distribution in the network. Scale-free networks have a power law node degree distribution $p(x) = Ax^{-\alpha}$ where for most of the observed networks $1 \le \alpha \le 3.5$. This implies that there are a large number of nodes in the network with a few links, while a small number of nodes (known as hubs) have a great number of links and keep the network together. The work of Barabasi and Albert [10] has lead to many scale-free models that usually introduce random preferential attachment mechanism and allow generation of a network with a power law distribution and a small average path length. However, the main drawbacks of the Barabasi-Albert model are the low clustering coefficient and the lack of ability to capture the community structures in the network. Thus, we created the scale-free communities (SFcom) model presented in [4] that incorporates the small world phenomenon and the scale-free property in such a manner that the obtained network is actually created as a network of a given number of clusters, i.e. communities, which have a scale-free property and are afterwards interconnected in a way that allows for the network as a whole to maintain the scale-free property. In this way, the obtained network also has a large clustering coefficient. The model tends to give representation of the real life situations wherein each group has at least one 'group leader' that is represented as a hub in the network.

The two proposed social network models were made in order to observe how the community structures influence the behavior of the ad hoc network on both the application and the physical layer of the network. The two variations of creating different communities offer another interesting insight in the network behavior. The SWcom offer a small-world network made out of nodes with a common node degree, thus creating a network that will be balanced when considering the frequency of interactions between nodes inside the community. The heavy load (in terms of interactions) can be found only at the nodes that act as groups interlinks.



a) SWCOM - 4 communities

b) SFCOM – 4 communities

Fig. 1. Example social networks with 100 nodes created with the SWcom model (a) and the SFcom model (b) with 4 communities.

On the other hand, when considering the SFcom network, this egalitarianism is non existent. The distribution of links inside the groups is according to the power law distribution thus creating bottlenecks at each group leader (hub) in the network. At the same time, these hubs also have the duty to act as interconnection points for inter-group communication. For visual comparison, on Fig. 1 example networks obtained with both of the models are shown.

3 Community Based Communication and Mobility Patterns

Using the findings of the theory of complex networks, we created models that can be used to depict the physical connections between network devices, the routing of the data packets in a communication network, or the end-to-end communication between the network users. The main goal of the models is to realistically capture the social network of the ad hoc network users. These models are afterwards used to create traffic and movement generators that will mimic the way the ad hoc network is going to be used by its users in a real life situation. The models incorporate both the small world and the scale free properties while emphasizing the community structures which are constantly observed in the real social networks.

In order to realistically simulate the conditions under which an ad hoc network would be used in real life situations as a means for information sharing between groups of humans, we used our SWcom and SFcom model in combination with specialized custom made application layers for the widely used open source NS-2 network simulator [11].

Our custom application layers use the information read-in from the generated social network using the SWcom or SFcom algorithm. The obtained network defines the relationships between the ad hoc network users and their need for communication. In this way we define the social links (information links) for each node in the ad hoc network. We simulate each node of the ad hoc network as a different node of the underlying social network with its given links to other participants in the ad hoc network. Each node is allowed to communicate only with its defined information links (in this way we let the node send and receive data information only from its known ‘friends’). Whenever a node sends a new message using this application layer, the node randomly chooses a destination from the pool of known information links. In addition to the custom application layers, we created a traffic generator that generates traffic with a given offered load over the simulation time.

In order to be able to incorporate the community influence on the node physical positions and movement patterns, we also created another tool that serves as a mobility scenario generator which is a modified version of the community mobility generator [12]. The generator is modified so that it works as follows: first, it reads in the generated social network according to one of our models, secondly, using the Girvan-Newman modularity method [13] it finds the communities in the network and than assigns proportional parts of the simulation area to each community. The nodes are uniformly scattered in the appropriate community area. During the simulation the nodes are moving within the boundaries of their respective community while sharing a common goal.

4 Simulation Scenarios

In order to investigate the impact of the different community structures on the ad hoc network performances we conducted several series of simulations. All of the simulations were performed on the SeeGrid infrastructure [14]. In our scenarios we observe the total end-to-end throughput in the ad hoc network (total received data bits per second in the whole network) while varying the offered load from 0.1 Mbps up to 7 Mbps.

The ad hoc network consists of 100 nodes that are uniformly distributed in a square area of 1 km^2. The nodes are equipped with radios that use the IEEE 802.11 protocol, while the multihop routing is provided using the AODV routing protocol [15], and on the transport level we use UDP. We were studying the ad hoc network for several different cases of node mobility for our movement generator: static nodes, nodes with average speed of 1 m/s, 2 m/s and 5 m/s. We decided to observe the performances of a social network with 4 communities which provided us with a network that has clustering coefficient and average path length that are closest to the ones observed in real life [2]. The additional parameters of both of the models are chosen so that 85% of the links in the network are links within the communities, while the rest 15% of the links are inter-community links.

One of the goals of the paper is to observe the influence of the existing community structures in the network on the logical and on the physical level separately for both of the social communities' models. Thus, the simulation scenarios also offer the possibility to use the created SWcom or SFcom social (information) network only on the logical (application) layer L, while the nodes are scattered and moving randomly in the complete simulation area P=GeomRND; or only on the physical layer P in which case the nodes are moving together within their communities, but the communication is completely random L=RND; or on both, physical and logical layers L and P wherein the communication and the movement pattern are according to the community modeling of one of the proposed models.

5 Impact of Social Communities on Network Performances

The first set of results is focused on presenting the ad hoc network performances when the network is modeled using the small-world and the scale-free communities social networks on the logical and/or physical layer of the network. The results obtained when the communities are modeled according to the SWcom are shown on Fig. 2, while the results obtained when the communities are modeled according to the SFcom are shown on Fig. 3.

The results presented on Fig. 2 allow for a fair comparison of the network performances obtained the traditional way with random traffic and movement together with the results obtained when modeling communities in the network using the small-world communities model. It can be concluded that taking into account the existing communities on the physical layer does not greatly change the behavior of the network, while the impact of the communities structures on the logical level is more than slight. The network performances rise rapidly when the traffic is modeled according to the social ties of the users. This is even more evident when both layers are modeled taking into account the small-world communities that are overlapping (L=SWcom and P=SWcom). An overall impression is that the performances are saturating when the offered load reaches 1 Mbps, while the maximum throughput achieved is also 1 Mbps. This shows that 1 Mbps is a maximum throughput that can be achieved in a network with a balanced, that is, uniform, traffic between the nodes of the community. This uniform traffic is a result of the clustered small world behavior where all of the nodes that belong to one community have identical node degree and the distribution of source-destination pairs of nodes is uniform.

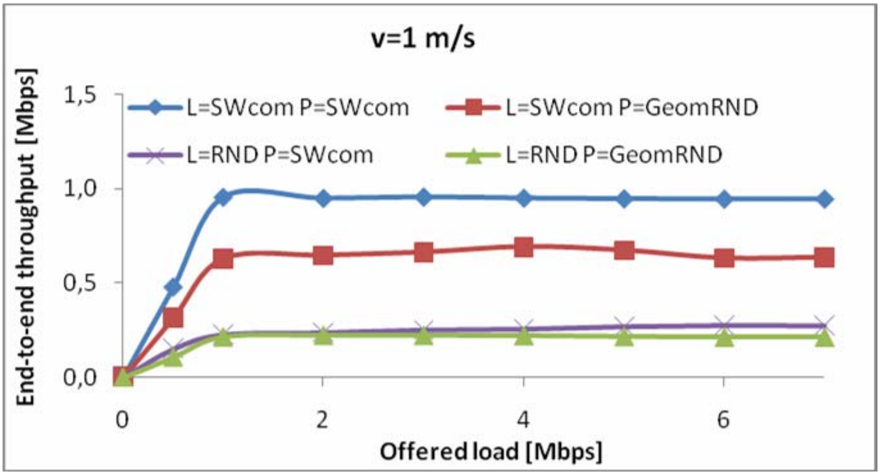


Fig. 2. End-to-end throughput in an ad hoc network modeled with and without small-world communities on logical and/or physical level using the custom SWcom application layer and movement generator for nodes that move with 1 m/s

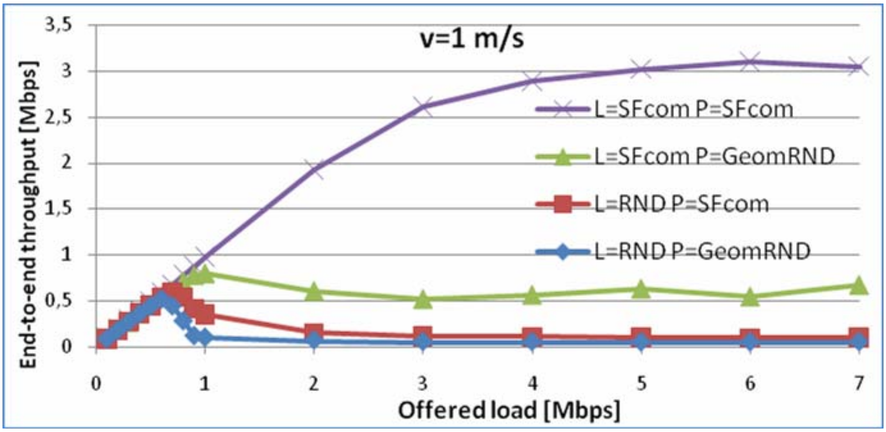


Fig. 3. End-to-end throughput in an ad hoc network modeled with and without scale-free communities on logical and/or physical level using the custom SFcom application layer and movement generator for nodes that move with 1 m/s

When the same analysis is done using the scale-free communities model there are some similar results, but also different trends in the obtained performances (see Fig. 3). The behavior of the network for communities only on the physical or on the logical level is very similar as it can be expected. However, when considering communities on both layers at once (please note that this case is the one to be found in real life situations), the network performances show different behavior. They reach more than 3 Mbps throughput for higher offered loads of 5 and 6 Mbps showing that the network does not get into saturation fast and allowing for a tremendous rising of the performances. This is somewhat peculiar while one is expecting to see lower performances because of the increased number of bottlenecks in this network model. However, care must be taken to consider that the network medium is shared and that in cases when in most of the communication the hubs are either source or destinations for the packet, it is much easier to solve the contention for the medium without the need for too many backoffs.

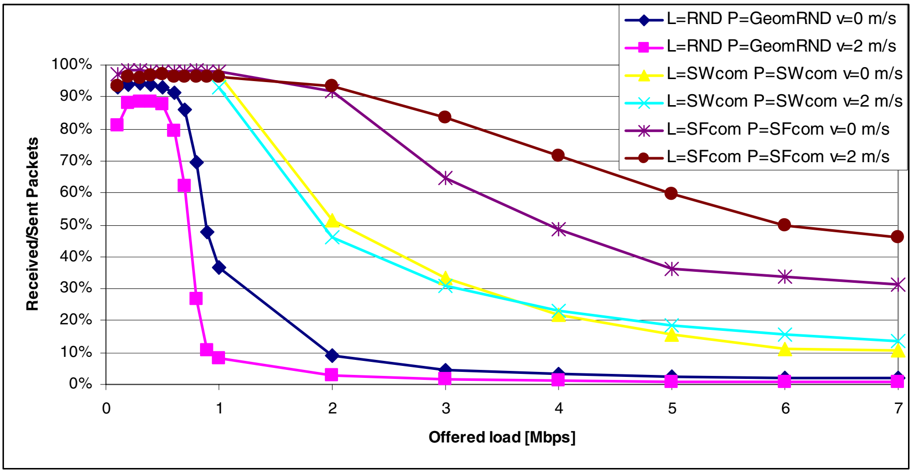


Fig. 4. Percentage of received over sent packets for network with and without communities modeled according to the SWcom or SFcom model for static and nodes that move with 2 m/s average speed

Fig. 4 offers some more insight on the ad hoc network performances from the perspective of how the percentage of received packets over the number of sent packets in the network changes when taking into account the social communities in the network using the small-world or the scale-free model instead of using the traditional random approach. It can easily be seen that while the random models predict only 1% of the packets to be delivered at their destinations for higher loads, when the traffic and mobility patterns depict the social activity of the network the simulations show that the percentage of received packets is several times higher than the expected and drops more slowly with the increasing load. Also, for the small-world model certain resilience to the node mobility is present, while in the scale-free communities case the node movement adds to the network performances increasing them for around 20% for higher loads. The increase of performances due to node mobility is one of the intrinsic characteristics of ad hoc networks and these sets of simulations just confirm this fact [5].

A comparison of the achieved network performances for various node speeds is presented on Fig. 5. The relative difference in the end-to-end throughput for the different models is clearly visible for static and nodes moving with 1, 2 or 5 m/s. The small-world communities modeled on both the logical and physical level show up to 10 times better performances than the traditional random scenarios. At the same time, the scale-free communities show improvements over the traditional models that are over 20 times higher. When comparing the two different communities types, one can conclude that the scale-free communities show outstanding performances which are in average 2 times better than the small-world communities. This is due to the distribution of the traffic inside the communities since in the both cases the number of community members as well as the number of communities is the same, which means that on the physical level, the modeling is done in the same manner since the mobility generator is concerned with these two parameters only.

The results clearly show that as much as the very existence of communities plays a major role in determining the network performances, the very distribution of the traffic in and outside these communities is a variable that has a vast impact on the network throughput.

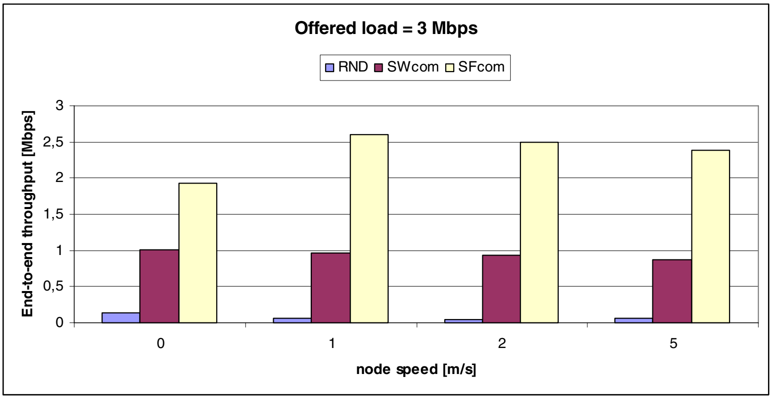


Fig. 5. Comparative analysis of the ad hoc network performances when the application and physical layers are modeled traditionally or with communities models for various node speed

6 Conclusion

In this paper we investigated the impact of the social communities structures that exist between the users of ad hoc networks on the actual network performances expected in real life situations. We argue that since the ad hoc network is going to be used as a tool to share information between a given group of people, the usage of the network in terms of traffic pattern and node mobility pattern is going to be governed by the rules of the social network established between the network users. Thus, care must be taken that while studying the network we work with scenario generators that will mimic this social behavior.

From this point of view, our main goal was to establish how the social grouping of users influences the overall throughput of the network. For this purpose we model the network users using small-world and scale-free communities on both the logical and the physical level. The results show that the communities on physical level cause a small rise in the network performances, while the communication pattern defined over communities on the logical level has a greater impact on the network performances.

Further more, when these two types of communities coincide, the ad hoc network performances are 10 to 20 times better when compared to the traditional simulation scenarios that are based on the random traffic and mobility patterns.

One of the main findings in this paper is that while the communities themselves have a great impact on the network performances, the way the traffic is distributed within the communities has an equal or even greater impact on the network behavior. Moreover, the results show that the power law distribution inside the community as well as in the social network as a whole creates focal points in the network traffic (the group leaders, or hubs in the network) which actually help to increase the network throughput by lowering the contention for the shared medium in the network.

References

1. Bakht, H.: Some applications of mobile ad-hoc networks. Computing Unplugged Magazine (2004)
2. Rheingold, H.: Smart Mobs: The Next Social Revolution. Macquarie University (2002)
3. Filiposka, S., Trajanov, D., Grnarov, A.: Analysis of small world phenomena and group mobility in ad hoc networks. In: CISSE 2006, USA (2006)
4. Fiiposka, S., Trajanov, D., Grnarov, A.: Performances of Scale-free Communities in Ad Hoc Networks, WirelessVITAE, Denmark (2009)
5. Kumar, B.R.A., Reddy, L.C., Hiremath, P.S.: Performance Comparison of Wireless Mobile Ad-Hoc Network Routing Protocols. Int. Journal of Computer Science and Network Security 8(6) (2008)
6. Lee, C.M., Pappas, V., Sahu, S., Seshan, S.: Impact of Coverage on the Performance of Wireless Ad Hoc Networks. In: Proceedings of the Second Annual Conference of the International Technology Alliance, UK (2008)
7. Shan, W., Jian-xin, W., Xu-dong, Z., Ji-bo, W.: Performance of anti-jamming ad hoc networks using directional beams with group mobility. In: IFIP International Conference on Wireless and Optical Communications Networks (2006)
8. Watts, D.J.: Six Degrees: The Science of a Connected Age. W.W. Norton & Company, New York (2003)
9. Watts, D.J., Strogatz, S.H.: Collective Dynamics of Small-World Networks. Nature 393 (1998)
10. Albert, R., Barabási, A.-L.: Statistical mechanics of complex networks. Rev. Mod. Phys. 74, 47–97 (2002)
11. The network simulator - NS-2, `http://www.isi.edu/nsnam/ns`
12. Musolesi, M., Mascolo, C.: A Community Based Mobility Model for Ad Hoc Network Research. In: Proceedings of the 2nd ACM/SIGMOBILE International Workshop REALMAN 2006, Italy (2006)
13. Clauset, A., Newman, M.E.J., Moore, C.: Finding Community Structure in Very Large Networks, arXiv:cond-mat/0408187v2 [cond-mat.stat-mech] (2004)
14. South Eastern European Grid Enabled Infrastructure Development, `http://www.see-grid.org/`
15. Perkins, C.E.: Ad hoc On-Demand Distance Vector (AODV) Routing Protocol, internet draft (2002)

Rule Induction of Physical-Chemical Water Property from Diatoms Community

Andreja Naumoski and Kosta Mitreski

Faculty of Electrical Engineering and Information Technologies,
Karpos 2 bb, P.O. Box 574, Skopje, R. Macedonia
{andrejna,komit}@feit.ukim.edu.mk

Abstract. In this paper we use the property of diatoms as bioindicators, to indentify which physical-chemical parameters are contained in the taken sample using machine learning algorithm – CN2. Important physical-chemical parameters such as conductivity, saturated oxygen, pH, organic chemical parameters and metals are important in the process of environmental monitoring. These physical-chemical parameters have influence on the entire lake web food chain, thus disturbing the organism's patterns and interactions between them, such as diatoms community. These communities have high coefficient of indication on certain process such as eutrophication and presence or absence of certain physical-chemical parameters, which means that they can be used as bio-indicators of water quality. The machine learning algorithm – CN2 can produce rules in a form IF-THEN which is suitable for organizing knowledge from diatoms abundance data. In literature the diatoms have ecological preference organized in the same manner. The experimental setup is build to satisfy not only the algorithm properties, but also the ecological knowledge of the diatoms community. We used several modifications of the algorithm, from which then we compare the compactness and coverage of the induced rule. Nevertheless, for regression problems we compare the correlation coefficient, root mean square error (RMSE) and relative root mean square error (RRMSE) or rule quality to point which experiment proved to be most accuracy and more general. Several of the rules are presented in this paper together with the evaluation performance.

Based on modifications of the CN2 algorithm parameters, we were able to extract certain knowledge form the data, which later have proved to be valid, or in some cases is novel for many newly discovered diatoms. In future we plan to investigate more modifications of the CN2 algorithm, also to implement multi-target rule induction and compare these results to the single target.

Keywords: CN2, coverage, water quality, diatoms, Lake Prespa, weighted.

1 Introduction

Rule induction is an area of machine learning in which formal rules are extracted from a set of observations. The rules extracted may represent a full scientific model of the data, or merely represent local patterns in the data.

The CN2 algorithm is a learning algorithm for rule induction. The original version of CN2 [1] used entropy as the heuristic evaluation function, while the later versions used accuracy as estimated by the Laplace relative frequency estimate [2] or the m-estimate [3, 4]. In the literature it can be found many improvements of the original CN2 method. The weighted approach, which is directly related to the rule learning method, is applied later in the paper on the diatom dataset.

In this paper we will use CN2 modification, to see how the algorithm performs on discovering rules from the diatoms abundance data of Lake Prespa. The focus of this paper is the investigation of the certain physical-chemical parameters that determinate the existence of certain diatoms, thus using diatoms to determinate the physic-chemical contents of the sample. This is done by using the measured data from the diatoms abundance and the machine learning technique.

Water quality is defined with the physical, chemical and biological characteristics of water. The most common standards used to assess water quality relate to drinking water, safety of human contact, and for health of lake, rivers ecosystems. These physical-chemical parameters are vital to understand how these systems function which in turn helps to identify the sources and fates of the ecological status of the lake.

The complexity of water quality as a subject is reflected in the many types of measurements of water quality indicators. Some of the simple measurements listed below can be made on-site; temperature, pH, dissolved oxygen, conductivity, and etc. More complex measurements that must be made in a lab setting require a water sample to be collected, preserved, and analyzed at another location. Making these complex measurements can be expensive. Beside these physical and chemical assessments of the water quality of lakes, there is also a biological assessment.

Bioindicators – biological assessments are species used to monitor the health of an environment or ecosystem. They are any biological species or group of species whose function, population, or status can be used to determine ecosystem or environmental integrity. In this paper we will used the diatoms community as bio-indicators collected from Lake Prespa to asset the physical-chemical parameters. There is known ecological preference for many of the diatoms, still the biologist discovery many new diatoms which have unknown ecological classification. Diatoms have narrow tolerance ranges for many environmental variables and respond rapidly to environmental change. This is making them ideal bio-indicators [7].

The data that we use were collected during the EU funded project TRABOREMA (FP6-INCO-CT-2004-509177). The measurements comprise several important parameters that reflect the physical, chemical and biological aspects of the water quality of the lake. The geographical location of the diatoms is not the limiting factor in the distribution of diatom species and the composition of communities; rather, the specific environmental variables prevailing at a particular location [8] are the limiting factors.

The remainder of this paper is organized as follows. In Section 2, we describe the machine learning methodology of CN2 algorithm that is used for knowledge discovery. Section 3 describes the data and section 4 explains the experimental design that was employed to analyze the data at hand. In Section 5, we present the obtained results in a form of rules and discuss them, while the section 6 concludes the paper.

2 Rule Learning for Knowledge Discovery

2.1 Rule Learning Basics

Predictive models can be written in more or less understandable forms. Among them, sets of 'if-then' rules are one of the most expressive and human readable model representations [9, 10]. Rule induction from examples has establish itself as a basic component of many machine learning (ML) systems, and has been the first ML technology to deliver commercially successful applications. The continued development of inductive techniques is thus valuable to pursue.

When compared to decision trees, rules are generally considered to be more understandable. Each rule, namely, represents an independent piece of knowledge that can be interpreted separately without other rules from the rule set while the decision rule has to be interpreted as a whole. A rule set consists of rules of the form 'IF condition THEN prediction; in addition, the rule set usually also has a default rule, which is used for prediction of examples that do not satisfy the condition of any other rule from the rule set.

There has been a lot of interest in rule learning within the machine learning and statistics communities. Most of the rule learning methods originates in the AQ series of methods [11, 12], which all employ the sequential covering algorithm. We will now briefly describe the covering algorithm as it is implemented in the CN2 method, which is one of more commonly used rule learning methods.

2.2 CN2 Algorithm

CN2 is an algorithm designed to induce "if...then..." rules in domains where there might be noise. The CN2 [1] is an algorithm that iteratively constructs classification (regression) rules.

The CN2 algorithm consist of two main procedures: a search algorithm performing a beam search for a good rule and a control algorithm for repeatedly executing the search. During the search procedure, CN2 must evaluate the rules if finds to decide which is the best. They are several metrics of rule quality is its accuracy on training data. In the original version of the CN2 algorithm the entropy is used, which behaves very similar to apparent accuracy, also prefers rules which cover examples of only one class. Later, more modifications of the algorithm are made.

2.3 Learning Regression Rules

So far, we have only discussed the learning of classification rules. Unfortunately, there are only a few approaches that can learn regression rules. A brief overview of these follows. [5] have developed a system, called SWAP1R, which transforms a regression problem into a classification problem. Later, the authors added the possibility of combining this method with the k-nearest neighbour method. The idea of transforming a regression problem into a classification one was further developed by [6].They developed a system called RECLA, which acts as a generic pre-processor and makes it possible to use an arbitrary classification method to solve regression problems. A rule learning system that learns regression rules directly was proposed by [13]. Another approach for learning regression rules is the patient rule induction method (PRIM) [14].

The methods presented so far learn rules directly. An alternative is to first learn a decision tree, and then convert it to a set of rules. The approach is applicable to both classification and regression problems. Rule sets constructed via trees may not be as understandable as rule sets constructed directly. A solution to this problem is post-processing of rules which results in smaller and more understandable rule sets [15].

2.4 Coverage and Compactness

Each rule should represent a general piece of knowledge. More generalization means that the rule covers more examples and in the end, it also means that the final rule set will have fewer rules and will be more comprehensible. Unfortunately, more generalization most often also means larger error in the model, and a compromise between the two must be found. The definition of relative coverage is straightforward. Sometimes, however, it is useful to introduce example weights that are not uniform. Each example e_i then has an associated weight w_i. The relative coverage of rule R_u in this case is simply the sum of weights of the examples covered by R_u divided by the sum of weights of all examples. For a given (partial) rule that covers a set of examples S, its quality is estimated as the average distance of an example in S to the prototype of S. We will call this the"compactness" of the rule.

We will use the compactness (average distance of an example covered by a rule to the prototype of this set of examples). The compactness takes into account both the attribute and target variable dimensions and is a weighted sum of the compactness along each of the dimensions (the latter are normalized to be between 0 and 1).

The compactness measure of a set of examples along one nominal attribute is calculated as the average distance of all examples in a set from its prototype. The distance between a rule r and a set of rules R can be defined as the average distance between the rule r, and each rule from the set R. For the attribute with K possible values (v_1 to v_K) the prototype is of the form ($f_1, f_2, \ldots, f_K$). The distance of an example with a value v_K from this prototype is equal to ($1-f_K$). The compactness measure for one numeric attribute is equal to the mean absolute error of the attributes value and it's mean. The values of numeric attributes are normalized in advance.

3 Data Description

3.1 Measured Data

The data that we have at hand were measured during the EU project TRABOREMA. The measurements cover one and a half year period (from March 2005 till September 2006). Samples for analysis were taken from the surface water of the lake at several locations near the mout of the major tributaries. In total, 275 water samples were available, 218 from the lake measurements and 57 from the tributaries. On these water samples both physico-chemical and biological analyses were performed. The physico-chemical properties of the samples provided the environmental variables for the habitat models, while the biological samples provided information on the relative abundance of the studied diatoms.

The following physico-chemical properties of the water samples were measured: temperature (°C), saturated oxygen (in %), Secchi depth (meter), conductivity (µS/cm), pH, nitrogen compounds (NO_2, NO_3, NH_4, inorganic nitrogen), total nitrogen (all in mg/L), SO_4 (mg/L), Total Phosphorus (mg/L), Sodium (Na) (mg/L), Potassium (K) (mg/L), Magnesium (Mg) (mg/L), Copper (Cu) (mg/L), Manganese (Mn) (mg/L) and Zinc (Zn) (mg/L) content.

The biological variables were actually the relative abundances of 116 different diatom species. Diatom cells were collected with a planktonic net or as an attached growth on submerged objects (plants, rocks or sand and mud). This is the usual approach in studies for environmental monitoring and screening of the diatom abundance [17]. The complete diatoms acronyms can be found in [18]. The sample, afterwards, is preserved and the cell content is cleaned. The sample is examined with a microscope, and the diatom species and abundance in the sample is obtained by counting of 200 cells per sample. The specific species abundance is then given as a percent of the total diatom count per sampling site [16].

3.2 Experimental Design

In this paper we have performed several analyses along several different scenarios taking the raw, as measured data from the lake ecosystem. These scenarios were applied on the lake measurements dataset, even the dataset which was gather from the TRABOREMA consisted from lake and river measurements. The entire dataset consist from 116 diatoms and the physical-chemical properties of the water describe earlier.

In the CN2 algorithm we have applied different parameter strategies; the standard covering method and weighted covering method. The Beam search algorithm was set to its default values. Decision list was used for prediction method and the rules were always added, even the algorithm has options to define adding rules *If* and *If Better*. For evaluating the performance of the algorithm on the data, was used 10 folds cross-validation. To assess the quality of the learned rules of knowledge that we have gain from the data, we have compared the correlation coefficient and RMSE for all the modified CN2 algorithms. In the next section only the best rules for each dataset are given and description of the ecological preference.

4 Results Using CN2 Algorithm from Lake Measurements

4.1 Standard Covering Method

In this section we present the TOP10 best rules according the values of the coverage and compactness product evaluation criteria. The rules were induced only for the lake measurement datasets. Many of the rules are small but not general, and some of these rules are large in length and cover more general knowledge.

In Table 1 are presented the experiment results from the coverage and compactness values of the given datasets using standard covering method. Ranking is done according the product between compactness and coverage. For example the *Rule 63* produced by the algorithm depicts the physical-chemical structure of the taken water sample with highest value. We can easily note that this rule compared with the *Rule*

59 represent very different combination of parameters. This conclusion lead that the existence of the diatoms is determinate by the physical-chemical structure of the water, which is true according the ecological reference found in the literature [8].

Table 1. Performance evaluation of Compactness and Coverage values for standard method

Rule No.	Covering Method = Standard			
	Number of Rule Produce	**Coverage* Compactness**	**Number of Rule Produce**	**Coverage**
1	Rule 63	6.40	Rule 63	131
2	Rule 62	3.48	Rule 62	87
3	Rule 58	3.47	Rule 59	73
4	Rule 59	3.02	Rule 58	55
5	Rule 61	2.20	Rule 60	40

Based on this conclusion, we can classify the diatoms in which water quality container belong based on several important physical-chemical characteristics of the water.

1. *Rule 63*: IF (*Navicula krsticii* (NKRS)) <= 0 THEN

Temp	SatO	SD	Conduc	pH	NO_2	NO_3	NH_4	TotalN
9.22	68.94	3.132	223.03	7.86	0.01	33.89	0.3	3.87
OrgN	**SO_4**	**TotalP**	**Na**	**K**	**Mg**	**Cu**	**Mn**	**Zn**
2.74	30.31	28.64	5.48	1.56	6.29	4.59	16.59	6.46

According the *Rule 63* the NKRS diatom can exist in mile water temperature, high values of conductivity, low levels of nitrogen components (NO_2) and metals (Mn and Mg), which in many cases are toxic. Using the rule induction to discover the diatoms ecological preference could lead to find new knowledge from measured data. The variations of the elements inside of the output rule vector for the CN2 algorithm must be found in certain boundaries of known physical chemical parameters of the water body. In this way the algorithm is in direct connection between the physical environment and the diatoms. The *Rule 59* depicts the environmental conditions where existence of the STPNN and GDEC diatoms in the water sample with high temperature and pH values, low values of NO_2 and NH_4 is suitable.

2. *Rule 59*: IF (*Staurosirella pinnata* (STPNN)) > 0 AND (*Geissleria decussis* (GDEC) <= 1) THEN

Temp	SatO	SD	Conduc	pH	NO_2	NO_3	NH_4	TotalN
19.37	87.11	2.65	187.33	8.25	0	1.4	0.3	2.56
OrgN	**SO_4**	**TotalP**	**Na**	**K**	**Mg**	**Cu**	**Mn**	**Zn**
2	32.19	22.9	4.4	1.79	6.45	5.03	8.32	10.72

The rule generated can be consisted from several AND inequalities of the test condition. This knowledge discovery leads to description of the appropriate environment for specific diatoms in diatoms community.

4.2 Weighted Covering Method

In this section we present the TOP10 best rules according the values of the coverage and compactness product evaluation criteria using the weighted approach. Table 2 present the evaluation data for the lake measurements datasets using the weighted covering method. Ranking according the product between compactness and coverage, for example the *Rule 4* produced by the algorithm depicts the physical-chemical structure of the taken water sample with highest value.

3. *Rule 4*: IF (*Achnanthes sp.* (ACH) > 0 AND *Amphora aequalis* (AAEQ) <= 0) THEN

Temp	**SatO**	**SD**	**Conduc**	**pH**	**NO_2**	**NO_3**	**NH_4**	**TotalN**
14.33	86.72	3.63	212	7.97	0.01	0.8	0.27	1.56
OrgN	**SO_4**	**TotalP**	**Na**	**K**	**Mg**	**Cu**	**Mn**	**Zn**
1.17	19	21.94	3.96	2.19	5.9	5.53	7.71	9.87

Very low test scores were gain using the weighted approach compared with the standard covering approach on the lake data. Nevertheless, the induced rules some of them were in line with the known ecological diatoms preference.

Table 2. Performance evaluation of Compactness and Coverage values for weighted method

Rule No.	**Coverage Method = Weighed**			
	Number of Rule Produce	**Coverage* Compactness**	**Number of Rule Produce**	**Coverage**
1	Rule 4	0.70	Rule 9	4
2	Rule 9	0.54	Rule 8	4
3	Rule 7	0.42	Rule 6	3
4	Rule 1	0.32	Rule 7	3
5	Rule 8	0.12	Rule 10	3

For example, *Rule 9* depicts the presents of two diatoms *Sellaphora pupula* (SPUP) and *Cyclotella juriljii* (CJUR) within the physical-chemical structure given by the rule, were the high levels of SatO means that this diatom is dependent from oxygen levels. Low levels of SD (Secchi Disk) indicated that the both diatoms are living in clear water, with moderate high values of pH.

4. *Rule 9*: IF (SPUP > 4 AND CJUR <= 2) THEN

Temp	**SatO**	**SD**	**Conduc**	**pH**	**NO_2**	**NO_3**	**NH_4**	**TotalN**
22.03	100.62	2.33	193	8.04	0.01	3.07	0.34	3.88
OrgN	**SO_4**	**TotalP**	**Na**	**K**	**Mg**	**Cu**	**Mn**	**Zn**
2.92	46.68	19.8	3.41	1.29	5.75	4	12.15	5.25

Compared with some of the rules generated from the algorithm for example *Rule 4*, it is easy and immediately can be noticed that the physical-chemical structure of the water is very different.

5. Default Rule:

Temp	**SatO**	**SD**	**Conduc**	**pH**	**NO_2**	**NO_3**	**NH_4**	**TotalN**
14.4	67.47	3.46	212	8.39	0.23	3.94	0.52	2.63
OrgN	**SO_4**	**TotalP**	**Na**	**K**	**Mg**	**Cu**	**Mn**	**Zn**
1.27	28.96	15.86	4.25	1.32	5.27	3.09	2.12	1.63

All the rules that have been gain from the rule induction process are compared with the default rule generated from the lake dataset. The Table 3 concludes our research based on the rule induction from the diatoms abundance in Lake Prespa. The regression performance indices lower performance from the standard covering method approach.

Table 3. Performance evaluation for TOP10 rules of the CN2 standard covering method based on the lake diatoms abundance dataset

Parameters	**Covering Method = Standard / Weighted**			
	CC - Train	**CC - Test**	**RMSE - Train**	**RMSE - Test**
Temp	0.71/0.19	0.32/0.05	4.64/6.47	6.75/6.63
SatO	0.51/0.19	0.21/0.08	16.08/18.37	19.32/18.74
SD	0.68/0.26	0.04/0.02	0.52/0.68	0.82/0.71
Conduc	0.61/0.13	0.29/0.04	21.9/27.557	28.31/27.94
pH	0.59/0.23	0.05/-0.13	0.5/0.621	0.73/0.66
NO_2	0.82/0.18	0.16/-0.03	0./0.0403	0.05/0.05
NO_3	0.77/0.17	0.28/0.04	1.35/2.09	2.37/2.14
NH_4	0.69/0.29	0.04/0.07	0.13/0.17	0.21/0.18
TotalN	0.68/0.24	0.21/-0.01	0.93/1.23	1.38/1.30
OrgN	0.62/0.23	0.06/-0.03	0.86/1.07	1.24/1.13
SO_4	0.61/0.14	0.02/0.00	18.24/22.68	27.48/23.24
TotalP	0.59/0.19	0.18/-0.04	12.35/15.00	16.13/15.92
Na	0.61/0.20	0.11/-0.04	1.66/2.05	2.30/2.15
K	0.65/0.34	-0.04/-0.13	0.50/0.62	0.79/0.70
Mg	0.67/0.18	0.12/0.04	2.09/2.78	3.09/2.88
Cu	0.58/0.42	0.06/-0.04	2.27/2.53	3.25/2.93
Mn	0.31/0.08	0.03/-0.05	15.92/16.7	17.73/16.85
Zn	0.55/0.20	0.16/-0.07	3.68/4.32	4.74/4.54

5 Conclusion

In this paper, we applied machine learning methodology, in particular CN2 rule induction algorithm to induce knowledge from diatoms abundance for taken measured sample water physical-chemical property in Lake Prespa. We have made several experiments from the diatoms community that has different settings and different environmental preferences using modifications of the CN2 machine algorithm.

The CN2 modifications, based on the different coverage method; standard and weighted produce rules that later have been compare with their correlation coefficient, RMSE and RRMSE, together with coverage and compactness ration. From the result of the experiment the standard coverage methods vs. weighted method have been proven to be more general in the knowledge expression and more accurate.

From the observations of the obtained rules for define physical-chemical vector values we can note that the given species could be found to exist in combination with other diatoms in certain water sample with different combination of parameters. With this kind of knowledge representation, we can compare this knowledge with existing biological expertise. The known ecological preference for the diatoms can be compared with the induced knowledge from the data. Diatom combinations also play very important role in direction of classifying their ecological preference.

Criticism for these models can be made, that the model does not utilize the whole diatom community, but rather a several number of dominant diatoms appear in the models, contrary to the established expert practice in the environmental monitoring where the whole community is evaluated regarding the water quality status. Nevertheless, TRABOREMA measurements are based on the work done for the saprobity and trophy status of Lake Prespa [3] and are aimed in detecting the principle physico-chemical parameters that are influencing the diatom communities in the region, and also their mutual correlations.

Building this kind of expert system that consists from modified rule induction algorithm leads to improvement of decision-making systems for environmental engineers. With this conclusion on mind, we plan to investigate more novel algorithms and possible CN2 modification in future. One direction is to modify the weighted algorithm, and adopt for multi-target algorithm procedure on different water quality class define by certain physical-chemical parameters.

References

1. Clark, P., Niblett, T.: The CN2 induction algorithm. Machine Learning 3(4), 261–283 (1989)
2. Clark, P., Boswell, R.: Rule induction with CN2: Some recent improvements. In: Kodratoff, Y. (ed.) EWSL 1991. LNCS, vol. 482, pp. 151–163. Springer, Heidelberg (1991)
3. Cestnik, B.: Estimating probabilities: A crucial task in machine learning. In: Aiello, L. (ed.) Proceedings of the Ninth European Conference on Artificial Intelligence (ECAI 1990), London, UK/Boston, MA, USA, Pitman, pp. 147–149 (1990)
4. Džeroski, S., Cestnik, B., Petrovski, I.: Using the m-estimate in rule induction. Journal of Computing and Information Technology 1(1), 37–46 (1993)

5. Weiss, S.M., Indurkhya, N.: Rule-based regression. In: Bajcsy, R. (ed.) Proceedings of the 13th International Joint Conference on Artificial Intelligence (IJCAI 1993), Chambéry, France, August 28-September 3, pp. 1072–1078. Morgan Kaufmann, San Francisco (1993)
6. Torgo, L., Gama, J.: Regression by classification. In: Borges, D., Kaestner, C. (eds.) SBIA 1996. LNCS (LNAI), vol. 1159, pp. 51–60. Springer, Heidelberg (1996)
7. Reid, M.A., Tibby, J.C., Penny, D., Gell, P.A.: The use of diatoms to assess past and present water quality. Australian Journal of Ecology 20(1), 57–64 (1995)
8. Gold, C., Feurtet-Mazel, A., Coste, M., Boudou, A.: Field transfer of periphytic diatom communities to assess short term structural effects of metals (Cd, Zn) in rivers. Water Research 36, 3654–3664 (2002)
9. Flach, P., Lavrać, N.: Rule induction. In: Berthold, M.R., Hand, D.J. (eds.) Intelligent Data Analysis, 2nd edn., pp. 229–267. Springer, Berlin (2003)
10. Mitchell, T.: Machine Learning. McGraw-Hill, New York (1997)
11. Michalski, R.S.: On the quasi-minimal solution of the general covering problem. In: Proceedings of the Fifth International Symposium on Information Processing (FCIP 1969), Bled, Yugoslavia. Switching Circuits, vol. A3, pp. 125–128 (1969)
12. Michalski, R.S., Mozetic, I., Hong, J., Lavrač, N.: The multi-purpose incremental learning system AQ15 and its testing application to three medical domains. In: Proceedings of the Fifth National Conference on Artificial Intelligence (AAAI 1986), Philadelphia, PA, USA, pp. 1041–1047. Morgan Kaufmann, San Francisco (1986)
13. Torgo, L.: Data fitting with rule-based regression. In: Žižika, J., Brazdil, P. (eds.) Proceedings of the Second International Workshop on Artificial Intelligence Techniques (AIT 1995), Brno, Czech Republic (1995)
14. Friedman, J.H., Fisher, N.I.: Bump hunting in high-dimensional data. Statistics and Computing 9(2), 123–143 (1999)
15. Friedman, J.H., Popescu, B.E.: Predictive learning via rule ensembles. Technical report, Stanford University, Stanford, CA, USA (2005)
16. TRABOREMA Project WP3, EC FP6-INCO project no. INCO-CT-2004-509177 (2005-2007)
17. WFD Water Quality - Sampling - Part 2: Guidance on sampling techniques (ISO 5667-2:1991) (1993)
18. Levkov, Z., Krstič, S., Metzeltin, D., Nakov, T.: Diatoms of Lakes Prespa and Ohrid (Macedonia). Iconographia Diatomologica 16, 603 (2006)

Semantic Interaction in Enterprise Data-Flow Visualization Environments: An Exploratory Study

Alberto Morell Pérez[1], Jorge Marx Gómez[2], and Carlos Pérez Risquet[1]

[1] Central University of Las Villas,
Faculty of Mathematics, Physics and Computing, Department of Computing Science
Carr. a Camajuaní, Km. 5, Santa Clara, Cuba
{amorellp,cperez}@uclv.edu.cu
[2] Carl von Ossietzky University Oldenburg,
Department of Computing Science, Business Information Systems I / VLBA
Ammerländer Heerstrasse 114-118, 26129 Oldenburg, Germany
jorge.marx.gomez@wi-ol.de

Abstract. Semantic interaction consists in interacting with the data by means of the image that represents it. In this paper, we analyze the possibility to add semantic interaction to a data-flow oriented visualization applications used in enterprise environments. For this purpose, we discuss a case study from the perspective of the type of interaction supported. We also review recent innovative approaches that attempt to use ontologies to link representation with meaning, using it to enhance user interaction and comprehension.

Keywords: Visualization, Semantic Interaction, Visualization Taxonomy, Ontology.

1 Introduction

Visualization is used as a tool to understand the data, to see the unseen. In this process, interaction plays a fundamental role, as it can reveal insights that a single representation or animation does not. One of the most interesting but less studied type of interaction is the semantic one, which consists in interacting with the data by means of the image that represent it. In this case, the user could be interested in the original data that is behind a single pixel or in the set of data that generates an area of the image. The purpose of data identification and selection is not only to get a detailed and specific view of that data subset but also to extract data for remodeling and further calculations [1].

Semantic interaction provides several advantages. First of all, it is more intuitive, as the user interacts with the final image but can think in terms of original values and context [2]. Furthermore, it allows the user to concentrate on the task at hand, instead of focusing in the interface itself, avoiding disruptive switch of the cognitive context. This makes the interface more "transparent", resulting in the enhancement of the cognitive productivity. Finally, it simplifies the coupling of input and output in the interface, suggested by Ware [3] as a mimesis of the objects of the real world, allowing that a visual object can potentially provide output as a representation of data and also potentially receive input.

However, there are major obstacles to achieve semantic interaction, especially in data-flow oriented systems, as one needs to be able to reverse the transformation of every module of the visualization pipeline. In the case of the enterprise environment, there is another complication, given by the type of data used, which is usually abstract and high-dimensional, without a pre-determined geometric representation.

In this paper, we study a data-flow oriented visualization application used in enterprise environments from the perspective of the type of interaction supported (Section 4). For this purpose, we propose a taxonomy which best adapt to our goals, the selection of which is discussed in Section 3. Then, we analyze the possibility to add semantic interaction (Section 5). In this section we also review recent innovative approaches that attempt to use ontologies to link representation with meaning, using it to enhance user interaction and comprehension, and show an example of the type of interaction envisioned. Finally, conclusions and issues for future research are given in Section 6.

2 Background

Interaction plays a fundamental role in the achievement of the traditional goal of visualization, the extraction of knowledge from data. A user's perception and understanding of a given visualization can greatly increase by its manipulation, e.g., changing one or more parameters at a time while observing the result, or simply moving the viewpoint around to reveal structure in the data that may be otherwise masked [4]. Besides, visualization is, by nature, an exploratory process: the perceived insight could raise new questions and hypothesis, which in turn trigger new interactions, in an iterative way [5,6].

In a general sense, visualization involves the transformation of raw data into an image, according to some specification, which includes the algorithm(s) to be applied and its parameters [6]. Following this basic model, there are two possible forms of interaction: with the transformation or with the data. In the first one, known as configuration and control, the user changes the algorithm or its parameters, to generate different images, according to the task at hand. In the second one, known as semantic interaction, the user identifies the application data by interacting with the visual information presented to him. The purpose of data identification and selection could be to get a detailed and specific view of a data subset or to extract data for remodeling and further calculations [1].

Usually, the transformation of the data into an image involves not one but several steps, each of which acts on some data to produce a new set of data, forming a chain or pipeline. These steps can be classified, depending on the type of operation they perform; in filter, if it produces a reduced subset of data; mapping, if an abstract geometrical object representing the data is generated; and rendering, if produces an image[1]. In visualization, this pipeline is commonly referred to as the data-flow network or visualization pipeline, and the steps as modules. One of the most desirable feature of a visualization system, is the possibility to build such a pipeline by the dynamic allocation of the available modules. These systems, which we will call Data-flow Visualization Environments (DVE), are more flexible, since they allow the user to customize a flow of visualization analysis. Another beneficial feature, in certain way related with the aforementioned, is the capability of the system to be extended with the incorporation, by the user, of new modules. AVS, Iris Explorer and Open DX are examples of this type of systems[2].

However, there are major obstacles to achieve semantic interaction in VDEs, as one needs to be able to reverse the transformation of the data as it passes from its original source, to the image. As the process involves the successive transformation of data in order to create the final rendered image, valuable information may be lost at each intermediate stage and the generated data first entering the visualization process may already have been stripped of much of its semantic content, specially if systems are aimed at satisfying general requirements [9]. Downstream modules have no possibility to access the information available further upstream. Besides, upstream modules store the original data but have no information about what happens further down the pipe and how the resulting image is generated[3].

To overcome this difficulty a variety of approaches are proposed in the literature. One of them is the Visualization Input Pipeline technique described in [1]. Here a backwards pipeline is constructed parallel to the original visualization counterpart. Each backwards-oriented module 'knows' the functionality and the parameters of the original module and inverses its data modification. Some application builders provide a similar feature. AVS and IRIS Explorer, for example, both use an augmented data-flow model supporting image probing facilities based on feedback loops – functions that accept a geometrical position, query the input data and return a value interpolated at the required point; but information of higher semantic level cannot be regained.

3 Methods

In order to study the possibility to implement semantic interaction in data-flow oriented visualization systems used in enterprise environments, we took on the task of analyzing the kinds of interactions available in these systems. Taking into consideration that it is very difficult to examine all existing systems and techniques, and in order to simplify the process of decomposing and organizing the

[1] This classification corresponds with the Upson et al. [7] model of Scientific Visualization. Similar steps of data transformation, visual structure generation, and view rendering is described by Card et al. [8] for Information Visualization.

[2] Other classification for these systems is Modular Visualization Environments (MVE).

[3] This depends on the architecture of the system.

problem space, we began our research by reviewing existing taxonomies of interaction techniques for visualization, and selecting one appropriated for our analytical purpose. After that, we explore a visualization system used frequently in enterprise environments. We tried that this system were representative in its area of application. We also reviewed articles describing visualization and interaction techniques in Information and Scientific Visualization (e.g., [10,11,12,13,14]).

We found that there are many taxonomic reviews relevant to visualization interaction techniques, but they have significantly different levels of granularity. Some categorize low-level interaction techniques, and other high-level user tasks. Some are oriented toward a specific application area (e.g. decision support environments [15] and graph visualization [12]) or research field (e.g. Scientific Visualization and Information Visualization) and others use a more holistic approach[4]. We agree with [16] in that former classification or surveys on visualization often mix up techniques and systems, specially in information visualization.

Taking into consideration these problems, we identified the following prerequisites the selected taxonomy must meet:

- to cover almost all the aspects and levels of the visualization process (e.g., data type, display mode, interaction style, analytic tasks),
- to include information and scientific visualization techniques,
- to analyze visualization and interaction techniques together, and
- to differentiate between visualization systems and techniques.

As a result of the study, we decided to use a taxonomy based in Qin et al.'s [16]. We will use the developer-oriented framework from this taxonomy, because we are more interested in the modification of the systems under analysis than in its use as a final user. According to this framework, each technique is classified along two dimensions: representation mode and interaction level. The representation mode dimension includes pixel-oriented, geometric projection, function transformation, icon-based, hierarchy-based, and graph-based representation. The dimension of interaction level includes manual, mechanized, and steerable interaction. We considered convenient to add the directness - Direct Manipulation (DM) or Indirect Manipulation (IM), taking into consideration the original work of Tweedie [19] on which the Qin et al. taxonomy is based.

Besides the analysis of the visualization techniques offered by the system under study, we also describe it along two categories: the scope and the architecture. The scope includes a general description and its main purpose; and the architecture, the general layout of the system (e.g., modular, monolithic, client-server).

4 Case Study

For the selection of the systems, we firstly chose a relevant application area in the enterprise environment. Then, a representative visualization system used in that area was selected. On the other hand, this area must also be characterized

[4] For more detailed comparisons of visualization taxonomies see [16,17,18].

by (1) the use of high data volumes, (2) the generation of complex - multivalued, multiparametric, temporal and spacial - data, and (3) the necessity of interaction with both the visualization and the raw data. Of course, we also take into consideration that the systems can be classified as a DVE. The area selected was Business Intelligence (BI) and the system, Pentaho.

Pentaho (http://www.pentaho.com) is a popular Open Source BI application suite made from free component applications. The Pentaho project was born in 2005 as the idea of BI veterans from Cognos, Hyperion, IBM, Lawson, Oracle and SAS.

Scope

The Pentaho BI Suite provides a full spectrum of business intelligence (BI) capabilities, including query and reporting, interactive analysis, dashboards, data integration/ETL and data mining. The suite uses a number of third-party Open Source components, as Mondrian OLAP Server and jPivot Analysis front-end, Firebird RDBMS, Shark and JaWE Workflow, Kettle EII and ETL, JBoss Application server, Hibernate and Portal, and Weka Data Mining.

Pentaho is a process-centric, solution-oriented platform. Every element of the system is handled as a process, and the solutions to the business tasks are modeled as a workflow of such processes, under the control of a Workflow Engine. Every step of each process is implemented as a standalone, re-usable component that can be directed to execute the activity required and can be used as a web service.

Architecture

The platform is implemented as a layered architecture (Fig. 1). At the core of the system is the solution engine, which loads and executes the BI processes (named action sequences). In order to do this, it runs the run-time engine, which in turn loads the needed resources and executes the action sequence. The run-time engine can be seen as a BI virtual machine, because it integrates all the BI components. Components are modules that can be added to the system and bring the BI features, such as reporting, charting, OLAP, ETL, email and workflow. Some of the components are self-contained, and other needs an external engine, e.g., an email or print engine. They form the lower layer of the system. At the other end, there is a user interface for interacting with the solution engine using a browser or desktop application, and an API, which can be accessed using several protocols such as HTTP, JMS, SOAP, AJAX, POJOs and BPEL.

This modular architecture is flexible and easily expandable with the use of plugins and extensions. The element that wire up all these components together is the action sequence interpreter. An Action Sequence is an XML document (with .xaction extension) that defines the smallest complete task that the solution engine can perform. It is executed by a very lightweight process flow engine and defines the order of execution of one or more component of the Pentaho BI Platform. Action sequences can be executed as part of a more complex

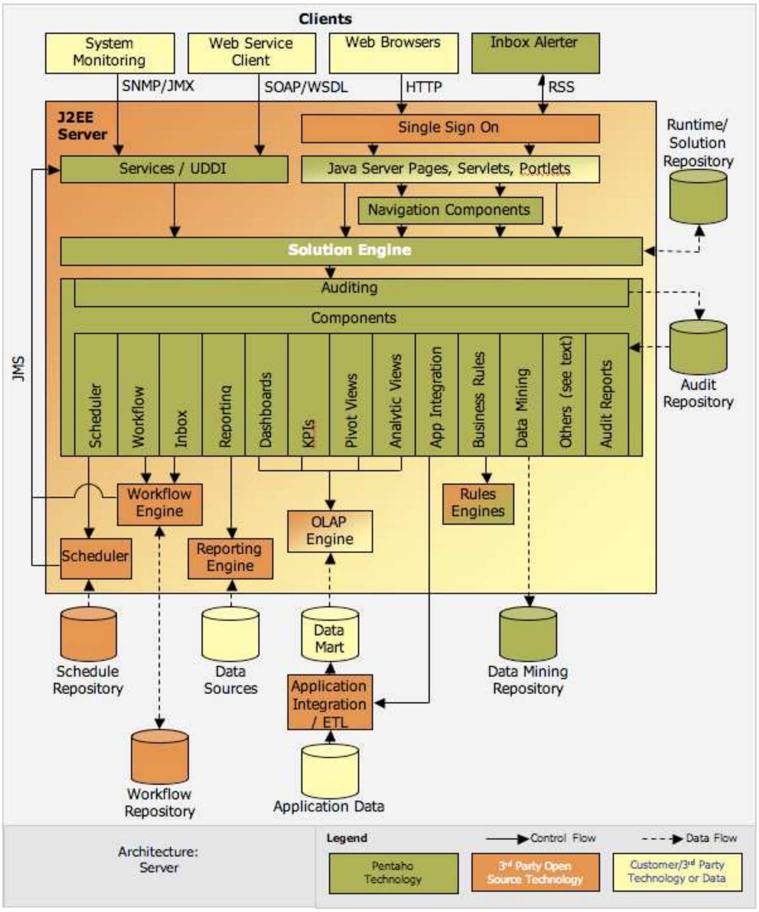


Fig. 1. Pentaho BI Platform components [20]

workflow, using the XML Process Definition Language (XPDL) and a XPDL-compliant workflow engine like Shark. In the case where a solution needs to be coordinated externally, any business flow defined in the system is available as web services and return their results via SOAP packages. This allows actions to be coordinated via an orchestration technology such a BPEL workflow engine or a remote application.

An action sequence defines the interaction between one or more actions. An action defines a task performed by a BI component. When the action sequence is executed, its actions are executed sequentially. Each action produces an output that the next action can use as input. This flow is conceptually similar to a pipeline, except for the fact that loops and conditions are permitted. There are another two BI components where similar data-flows can be defined, Pentaho Data Integration and Data Mining, but they are not included in this research.

Visualization techniques

Among the BI end-user capabilities available in the Pentaho Suite, the ones that made a more intensive use of graphics are reporting and dashboards. Reports are used to present more static information, with absence of interaction; for this reason we concentrated our analysis in dashboards[5] [21].

In Pentaho, a dashboard can contain a number of visual components, of which four are more suitable for the presentation and analysis of great amount of data: charts, dials, maps[6] and time plot. Among them, charts are the most varied in terms of available types. Some of the most representative visualization techniques employed to generate this graph are:

(A) Scatter plot
Representation: Cases are represented by locations of points (geometric projection)

Interactivity: (1) Users can select a single item by selecting any data point (manual DM), (2) data is hidden (mechanized DM) or filtered (mechanized IM) by selecting ranges.

(B) Pie plot, bar plot
Representation: A bar or a circle is used to represent the data (icon based)

Interactivity: (1) Users can select a single item by selecting the corresponding icon (e.g., bar or pie section)(manual DM), this can trigger an action specified in an url-template; (2) A tooltip may be displayed when mouse hovers over a data item (manual IM)

(C) Glyph
Representation: Cases are represented as complex symbols whose features are functions of the data (icon based). The symbol (parameter) depends on the chart type, e.g., circles (diameter) in bubble charts.

Interactivity: (1) Users can select a single item by selecting the corresponding symbol (manual DM), this can trigger an action specified in an url-template; (2) A tooltip may be displayed when mouse hovers over a data item (manual IM)

(D) Geographic map
Representation: Cases are represented by points located in a two-dimensional map (geometric projection), each point can encode information using a glyph (e.g., glyph shape and color)

Interactivity: (1) Users can select a single item by selecting any data point (manual handle), the details of each location are obtained with a Pentaho Action result content; (2) The user can navigate through the map (manual DM, mechanized IM)

[5] Using the Community Dashboard Framework (CDF) implementation.

[6] Actually, there are two map components, but they are functionally similar.

5 Discussion

Pentaho can not be considered a pure DVE, as MVEs are, because it does not allow the composition of several visualization techniques in order to obtain the final image. The visualization components, when used, are located at the end of the workflow. Furthermore, the number of visualization techniques available in the system are really small, in comparison with other more specialized tools, like GeoVista Studio or GraphViz. However, its workflow paradigm is very similar to the visualization pipeline, and its modular architecture allow the addition of new components, which would allow to settle these setbacks.

Regardless the use of semantic interaction in Pentaho, there is not a mechanism to obtain a data that is several steps back in the workflow. In the case of the dashboards, its elements can interact between themselves, but not with individual elements of an action sequence. Dashboards also allow drilling down and up, but only along the dimensions defined in the used cube.

Several alternatives emerge for the implementation of semantic interaction in Pentaho. At a first look, it seems logical to apply the approach taken by other DVEs, like AVS and IRIS Explorer. These systems provide probing modules, which evaluate a dataset at some discrete points or plot it along a user defined line segment. However, this is a partial solution, because information of higher semantic level can not be regained.

One of the most promising alternative is the use of ontologies. As we said at the beginning, the purpose of semantic interaction is data identification and selection. Due to the great amount of data that is generally visualized, it is unlikely that the user want to interact with a particular data instance. Normally the intention of interaction is an object with certain meaning, or semantic, like a fluid line or a cluster. However, the identification of such an object is not always trivial, specially when visualization is used for exploration, where the user has no idea what to look for. Ontologies offer a way to solve this problem, allowing to link the representation with its meaning.

Once identified the intended object in the final image, this information must go back in the pipeline, to determine the original data it represents. If we see the visualization pipeline as a sequence of transformation steps (filter, mapping, rendering), each one changing the semantic level of the input (from data to image), then the ontology-based mapping must be used at every step. Following this idea, Duke in [22] designed a component to provide ontology support to the visualization pipeline, and described its implementation in the Visualization Toolkit (VTK). In the same line of reasoning, Mikovec et al. [23] use semantic information in the development of methods for optimizing the visualization process. Here, the semantic metadata are used for data reduction and for emphasizing interesting parts of the data.

As a last remarks, even though it was not an explicit requisite, the open source condition of Pentaho was an important factor for its selection. It allows to know in details the implementation as well as to modify the code for testing new features.

6 Conclusions and Future Research

Semantic interaction is an important feature for data visualization, that provides additional means for an analyst to explore their data sets. In this paper, we evaluate the availability of this feature in a data-flow oriented visualization applications used in enterprise environments, namely, Pentaho.

In our preliminary study, we detect that most Information Systems and Data Analysis Tools used in enterprise environments are primarily output-oriented. Users can specify and change the parameters that are controlling the analytical process, which results in different images or data representations, but no mechanism is provided to really interact with the application data (semantic interaction), that has been changed step by step by the analytical process. This thesis was confirmed by our case study.

Of course, only one system is not enough for arriving to definitive conclusions. We propose to extent this research with the analysis of others systems using the methodology developed in Section 3.

This work also discussed alternatives to add semantic interaction in the systems under analysis. We considered that the use of ontologies to associate meaning to the output of every step of the visualization pipeline, is a promising way. This path has to be explore further.

The implementation of semantic interaction in data visualization applications used in enterprise environments has and additional problem, given by the type of data used in them, which is usually abstract and high-dimensional, without a pre-determined geometric representation. For this reason, it could be difficult for the user to identify the form of interaction with the features or categories represented in the image. To confront this problem, it is required the design of new visualization metaphors for the creation of visualization's meta-controls susceptible to semantic interaction for data analysis and decision making tasks.

References

1. Felger, W., Schroder, F.: The visualization input pipeline-enabling semantic interaction in scientific visualization. Computer Graphics Forum 11(3), C139–C151 (1992)
2. Schröder, F.: Ape—the original dataflow visualization environment. SIGGRAPH Comput. Graph. 29(2), 5–9 (1995)
3. Ware, C.: Information Visualization: Perception for Design. Morgan Kaufmann Publishers Inc., San Francisco (2004)
4. Zudilova-Seinstra, E., Adriaansen, T., van Liere, R.: Trends in Interactive Visualization: State-of-the-Art Survey. Springer Publishing Company, Incorporated, Heidelberg (2008)
5. Jankun-Kelly, T.: Visualizing Visualization: A Model and Framework for Visualization Exploration. PhD thesis, Univ. of California, Davis (2003)
6. van Wijk, J.J.: The value of visualization. In: IEEE Visualization Conference, p. 11 (2005)
7. Upson, C., Thomas Faulhaber, J., Kamins, D., Laidlaw, D.H., Schlegel, D., Vroom, J., Gurwitz, R., van Dam, A.: The application visualization system: A computational environment for scientific visualization. IEEE Comput. Graph. Appl. 9(4), 30–42 (1989)

8. Card, S.K., Mackinlay, J.D., Shneiderman, B.: Readings in information visualization: using vision to think. Morgan Kaufmann Publishers Inc., San Francisco (1999)
9. Chatzinikos, F., Wright, H.: Computational steering by direct image manipulation. In: VMV 2001: Proceedings of the Vision Modeling and Visualization Conference 2001, Aka GmbH, pp. 455–462 (2001)
10. Elvins, T.T.: A survey of algorithms for volume visualization. SIGGRAPH Comput. Graph. 26(3), 194–201 (1992)
11. Ma, K.L.: Image graphs—a novel approach to visual data exploration. In: VIS 1999: Proceedings of the conference on Visualization 1999, pp. 81–88. IEEE Computer Society Press, Los Alamitos (1999)
12. Herman, I., Melançon, G., Marshall, M.S.: Graph visualization and navigation in information visualization: A survey. IEEE Transactions on Visualization and Computer Graphics 6(1), 24–43 (2000)
13. Fekete, J.D., Plaisant, C.: Interactive information visualization of a million items. In: INFOVIS 2002: Proceedings of the IEEE Symposium on Information Visualization (InfoVis 2002), Washington, DC, USA, p. 117. IEEE Computer Society, Los Alamitos (2002)
14. Moere, A.V.: Beyond the tyranny of the pixel: Exploring the physicality of information visualization. In: IV 2008: Proceedings of the 2008 12th International Conference Information Visualisation, Washington, DC, USA, pp. 469–474. IEEE Computer Society, Los Alamitos (2008)
15. Adnan, W.A.W., Daud, N.G.N., Noor, N.L.M.: Expressive information visualization taxonomy for decision support environment. In: International Conference on Convergence Information Technology, vol. 1, pp. 88–93 (2008)
16. Qin, C., Zhou, C., Pei, T.: Taxonomy of visualization techniques and systems - concerns between users and developers are different. In: Asia GIS Conference 2003 (2003)
17. Yi, J.S., Kang, Y.a., Stasko, J., Jacko, J.: Toward a deeper understanding of the role of interaction in information visualization. IEEE Transactions on Visualization and Computer Graphics 13(6), 1224–1231 (2007)
18. Brodlie, K., Noor, N.M.: Visualization notations, models and taxonomies. In: Lim, I.S., Duce, D. (eds.) EG UK Theory and Practice of Computer Graphics, pp. 207–212 (2007)
19. Tweedie, L.: Characterizing interactive externalizations. In: CHI 1997: Proceedings of the SIGCHI conference on Human factors in computing systems, pp. 375–382. ACM, New York (1997)
20. Pentaho Corporation: Pentaho open source business intelligence platform white paper (2008), `http://www.pentaho.com/products/demos/osbi_technical_white_paper.php`
21. Pentaho Community: BI Server 2.x-3.x Community Documentation, `http://wiki.pentaho.com/display/ServerDoc2x/BI+Server+2.x-3.x+Community+Documentation` (2009)
22. Duke, D.J.: Linking representation with meaning. In: VIS 2004: Proceedings of the conference on Visualization 2004, Washington, DC, USA, p. 598.5. IEEE Computer Society, Los Alamitos (2004)
23. Mikovec, Z., Klima, M., Slavik, P.: Semantic driven visualization. CODATA Prague Workshop. Information Visualization, Presentation, and Design (2004)

A Beehive-Like Multi-agent Solution to Enhance Findability of Semantic Web Services and Facilitate Personalization within a P2P Network

Ammar Memari, Mahmoud Amer, and Jorge Marx Gómez

Business Informatics I, Very Large Business Applications,
Carl von Ossietzky University,
Ammerlaender Heerstr. 114-118
26129 Oldenburg, Germany
{memari,amer,marx-gomez}@wi-ol.de
http://vlba.wi-ol.de

Abstract. The market of Semantic Web Services is a heterogeneous volatile environment. Ability of the enterprise to adapt to it by finding relevant and high quality resources is crucial. This paper presents an approach for personalization in a peer to peer network by continuously searching for relevant resources using a beehive like mechanism and aggregating results in a fuzzy manner. We consider diversity, scrutability and efficient traversal as key features to face difficulties of constructing a successful market. And link the results and findings of our study to the quality of electronic services, and will show how the proposed solution shall enhance the overall user satisfaction based on specified dimensions used to measure its level.

Keywords: beehive, multi-agent system, mobile agent, personalization, peer to peer network, Semantic Web Services, search engine, Quality of Electronic Services, fuzzy logic.

1 Introduction and Background

A main benefit of Service Oriented Architecture (SOA) is giving the possibility to delegate development of some functionality to the service creator while allowing concentration on the full view of the system. Problems come with the ascending acceptance and adoption of this method in the form of Web Services. Since many vendors are and will be flooding the Web Service market every day. Finding our sought-after service thats buried under a big pile of useless irrelevant ones can be difficult and time consuming. Adding semantics to the description of Web Services can raise findability by allowing the repetitive task to be assigned to machines, and aids in loosening coupling and lowering dependency which can be achieved by using semantic relations to allow the requester to have a clearer idea about available alternatives. To solve the findability problem, IDEAS (Intelligent Discovery of Enterprise Architecture Services) concept was introduced as a step ahead in the field of intelligent discovery of Semantic Web Services [1], to take the burden off the shoulders of the human user, and delegate the task of discovery to intelligent machines. IDEAS resides within a Web Services market, a commercial P2P network. A distinctive feature of this market is high unpredictability; this feature originates from a number of facts:

1. No conditions are required to enter the market since it is a P2P network; that enables new players to join in any minute to provide their Web Services, and allows players of all sizes to co-exist in this market providing a wide range of quality and prices.
2. As an e-market, it is quickly affected by the slightest of changes, conforming to a new customer need, market trend, change in offer and demand is almost instantaneous.
3. Internet is the connection medium in this market, not only in the brokering (discovery) phase as in most of e-markets but also in the delivery (execution) phase. Variations in traffic over the internet can directly affect the quality of provided Web Services.
4. Considering compound Web Services, the change in market state caused by any of the aforementioned factors which affects a Web Service, also affects all possible combinations that include it. In addition to that we can add factors that shape traditional markets such as mergers and acquisitions.

Determining locations of the best resources within such a near-chaotic environment is a difficult time-consuming task. Adapting to the new circumstances of the environment as fast as possible is crucial. A beehive-like approach seems logical within such an environment, an approach that is originally influenced by the works of Thomas D. Seeley [2], and James Surowiecky [3] after him. The case of the environment we have in hand is comparable to the case of a beehive trying to find nectar sources.

In this context, we can benefit from the work accomplished in [4] and [5], Where authors have tried to come up with a model of a beehive, and to build a search engine based on it. Some practical statistics and results were achieved by observation of different situations, and more research is still being conducted in that area.

The rest of the paper will contain a description of the beehive model and its working mechanism, follows that a section arranged around two main problems and how to approach solutions based on that model. next we will show the impact of this model on different dimensions of electronic services' quality, and the last section is assigned to discuss expected misuse cases and their workarounds.

2 From the Beehive Model to the Mobile Agents-Based Solution

A beehive model was proposed by Pavol Navrat and Martin Kovacik [4], which is a modified model of the one proposed by Fabiana Lorenzi et. al. [6]. We will use a modified version of the Navrat and Kovacik model in our approach, main differences come from:

- Our model is for a more specific case, P2P network; semantically defined Web Services are the resources we're trying to search among. Whereas the other model is more generic and it used the term web search in a rather loose sense.
- Our case is an asynchronous search mechanism that preprocesses the search space in order to enhance results of conventional search by getting more relevant results. Search in our case is not initiated by a query received from the user, rather it is a continuous automatic process which will receive user criteria and preferences from a user model stored and maintained by the requester host.
- Rather than offline searching in a frozen part of the web, as in the case of the former model, our case involves searching in a living part where Web Services are in continuous change, old ones might disappear and new ones are added continuously. Searching in the open corpus is what gives this method its efficiency compared to other searching methods.
- The logic used by the bees in [4] to evaluate quality of a pasture, dance (advertise) for it, and decide whether to follow a dancing bee is a crisp logic, and no diversity in beliefs or methodology of the working bees was introduced; that can be the reason behind the experimental results found in the same reference where although a big number of bees (2500) was used, they nearly missed the best recommendation. Whereas our model uses a fuzzy logic approach and introduces diversity in beliefs and methodology.

3 Main Benefits of Agents' Action as a Beehive

3.1 Traversal of the P2P Network

Two main drawbacks in P2P networks are latency and incompleteness of traversed nodes. Latency comes from the huge number of peers which will receive the query very late in the process. Incompleteness comes as a result of the freedom of peers to get on and off line arbitrarily, rendering some other peers unreachable, and preventing the network from being completely traversed. These two drawbacks can be overpassed by using power peers, which play the role of central registry points giving the network multiple core points, but that compromises decentralization and constitute points of failure, manipulation and produces extra maintenance effort.

As suggested in IDEAS (see [1]) to trouble-shoot these two drawbacks, a local cashing registry (namely In-UDDI) will be maintained. This registry will contain potential services gathered from the web based on a user model by a software agent working continuously in the background to keep this cashing UDDI fresh and up-to-date. This mechanism increases probability of querying the majority of peers. To overcome latency, search queries are executed first on this registry with high probability to hit relevant results.

Traversing the P2P network in our approach is suggested to follow the honey bees navigation system as illustrated in Fig. 1. Mobile agents go in scouting surveys and come back to run on the home peer again, where they get to know about their next destination.

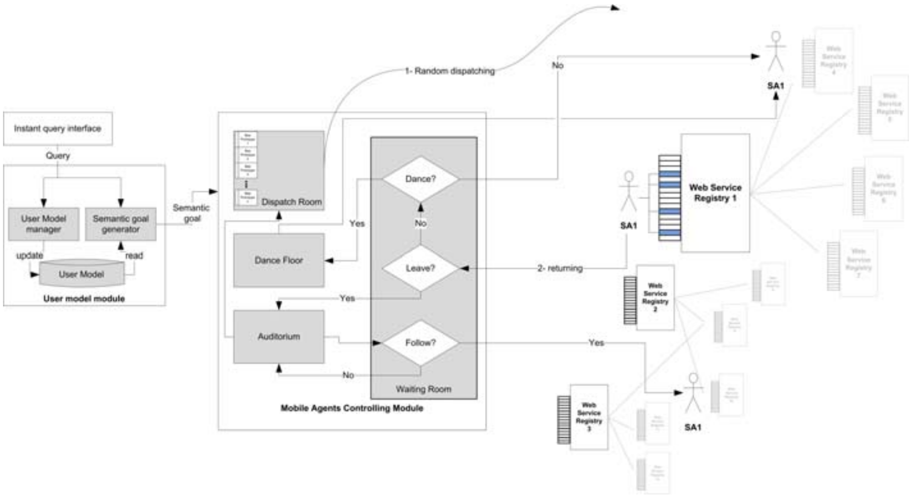


Fig. 1. The modified beehive model, showing an agent (SA1) as it's been dispatched randomly (to WS Registry 1) then either decides not to leave its source and migrates to a neighbor of it (WSR 4), or leaves its source and follows another agent to a neighbor of (WSR 2). On the left side we can see the user model module, which initiates searching activities triggered either directly through an instant user query or indirectly by generating semantic goals (semantic query) from the stored user model.

3.2 Maintaining Diversity to Solve Heterogeneity

A main point in the discovery and selection process is the matchmaking; which is the process where a match or no match is decided between requirements of the request and capabilities of the advertised service. Based on a number of criteria and different methodologies, needs of the requester are compared to a set of available services, and in many cases a list of services is returned, ordered by relevance or what is called Degree of Match (DoM). Different studies about Semantic Web Service matchmaking were conducted; they differ according to many aspects [7]:

- Whether to use ranking and DoM or just return either a match or no match.
- Matching parameters, functional and non-functional properties.
- Service composition support.
- Matchmaking algorithm type: logic-based, hybrid-based, similarity-matching e.g. [8] or graph-based matching e.g. [9].

In each application domain, these aforementioned aspects receive different weights, affecting the decision of which matching approach to use. No single matchmaking mechanism has proven to be significantly superior, and we can add to this that no single language for expressing Web Service semantics was approved as dominant.

Instead of struggling to find the one and only matchmaking method, we can keep the diversity among them and in the same time find a mechanism to aggregate their outcomes and produce collective findings. Having diversity among working bees allows us to avoid pitfalls mentioned earlier which are caused by too much consensus (see from the beehive model to the mobile agents-based solution). Working bees will be created by instantiating prototypes that hold diverse matching algorithms, and initialized with randomly generated beliefs. When these dissimilar bees come back to the hive, they will be holding different opinions about surveyed sources. These opinions are aggregated using the waggle dance method, in which a bee advertises her opinion on the dance floor, while other bees in the auditorium apply a fuzzy logic model that features uncertainty to sum up all opinions of the dancing bees into a decision upon which dancer to follow.

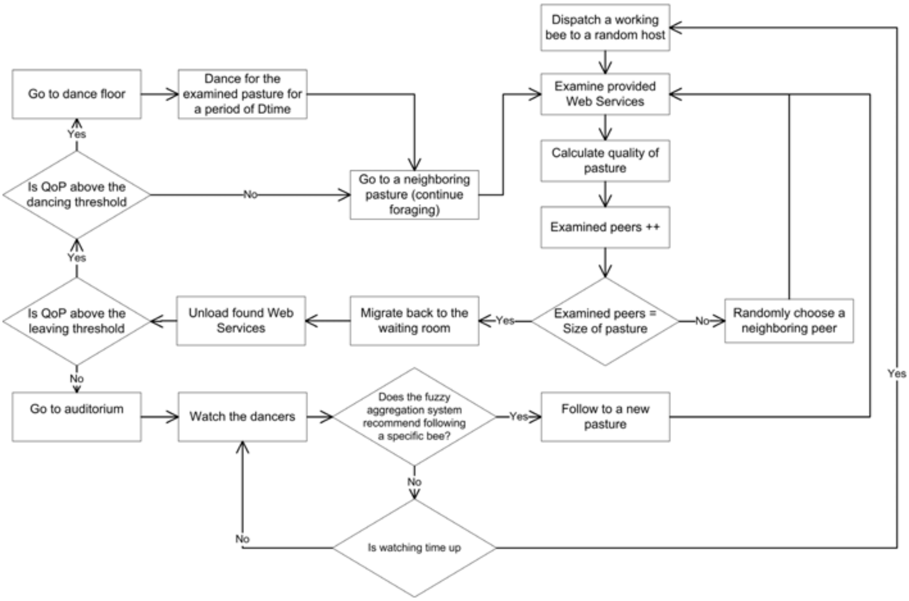

Dispatch a working bee to a random host
Go to dance floor
Dance for the examined pasture for a period of Dtime
Examine provided Web Services
Yes
Calculate quality of pasture
Is QoP above the dancing threshold
No
Go to a neighboring pasture (continue foraging)
Examined peers ++
Yes
Is QoP above the leaving threshold
Unload found Web Services
Migrate back to the waiting room
Yes
Examined peers = Size of pasture
No
Randomly choose a neighboring peer
No
Yes
Go to auditorium
Watch the dancers
Does the fuzzy aggregation system recommend following a specific bee?
Yes
Follow to a new pasture
No
No
Is watching time up

Fig. 2. Flowchart diagram depicting different decisions and states of a working bee

A source of Web Services is analogous to a pasture of flowers; a pasture in our case is defined as a set of peers grouped using one of these two relations:

1. Neighboring relation: all the peers are neighbors to at least one of them. This relation is based on the assumption of neighbor to neighbor coherency, i.e. on the fact that if a peer has services relevant to user's requirements, then there is a good possibility that its known peers also have relevant ones. This relation can be noticed in peer to peer file sharing networks.
2. Semantic relation: all the peers contain Web Services that are connected to one domain ontology. This relation requires a reverse resolving mechanism to be implemented in order to resolve all services connected to the domain ontology by having the ontology alone.

4 Benefits of the Proposed Model in Improving the Quality of Electronic Services

Electronic services are no longer regarded as trendy Internet applications; rather, customers have become more and more demanding. Also, they are less tolerant to poor services performance, and it is the delivery of high services quality that makes customers come back and buy again [10].

Based on the model of Amer and Marx-Gomez [11], we shall study the effect of using our proposed solution in enhancing the user perceived quality of service. Figure 3 shows the quality of electronic services dimensions, as we can see the core dimensions are: site features, security, responsiveness, reliability, accessibility, information, communication, personalization, delivery, ease of use, and customer support. For further information about the model please refer to [11].

Our proposed solution has the impact on the following dimensions:

1. **Privacy:** As mentioned earlier, the search is not initiated by a query received from the user, rather it is a continuous automatic process which will receive user criteria and preferences from a user model stored and maintained by the requester host. This means that instead of having the search profiles stored on a website or in the Internet, it is stored locally, which puts the user model under the user's control, and hence allowing for higher level of privacy provided to the user during the searches.

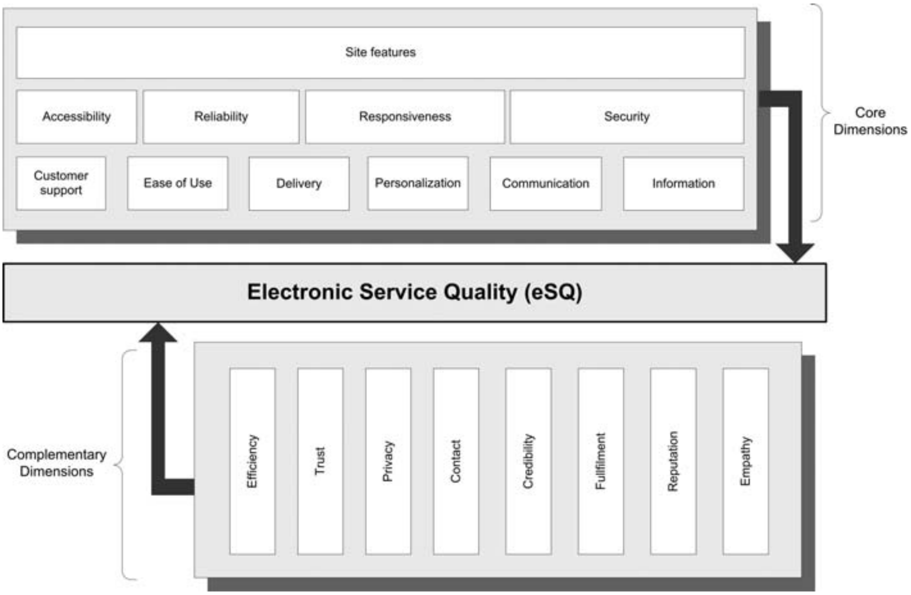


Fig. 3. Electronic Service Quality Model

2. **Personalization:** First, Using a continuous search-and-cash mechanism leads to creation of a personalized search space, this mechanism is based on semantic goals that are generated from a user model. This model is divided into smaller parts by the semantic goal generator (Fig. 1) and different bees might be assigned semantic goals of different parts of it. A similar issue was discussed also in the works of Pavol Navrat [4] and [5], by dividing search criteria into separate attributes, and assigning an attribute or more to a set of working bees. As the proposed application model is designed to generate queries based on the user model, it features higher flexibility. This flexibility and customization empowers the user to develop and maintain a customized user model based on his preferences and desired specifications. Giving the user these choices enhances the level of personalization to match different applications, skill levels, and the size of the search space needed in order to find the right solution. Second: not only the user has the choice to generate queries in the previously explained manner, but he also can develop a customized algorithm to adapt to his finest needs. Having the user model as an independent unit makes it easier for many enterprises to extend it to different application areas in different departments, and makes it possible for these enterprises to integrate it with currently existing systems and software to enhance and improve their operations. That allows moving to a new level of personalization, by permitting the user to personalize not only the static representation of his preferences (user model), but also the way these preferences are processed. Moreover, more than one matchmaking algorithm can be used simply by assigning different algorithms to different mobile agents as mentioned above, complying with the model suggested in [12].
3. **Accessibility:** Another technique used in the proposed solution is to store a local repository in order to cache the most relevant information as an extraction of the space search. Using this cached repository shall make it possible for the systems to search the most relevant information in the search space first, which shall not only save network bandwidth and improve response time, but also provide greater accessibility for the most relevant searches by the user. On the other hand, dealing with a commercial network of peers leads to a competition for higher positions on returned search result; in order to prevent manipulation of results by providers, the matching algorithm is implemented within the code of the mobile agent, and it is not only trusted, rather also customized and parameterized by the requester to satisfy his finest needs.
4. **Security:** The techniques used in the proposed solution yields to a higher level of trust and scrutability, because a predictable behavior is performed since the algorithm is hand-picked, and a clear explanation of the reasoning process can be generated upon request. Enhanced security for private information is achieved since the provider will have no direct access to the user's preferences. Instead, the user model is accessible only by the mobile agent using pre-specified and customized routines. Thus reducing phishing possibilities of private information.

5 Potential Misuse Cases and Workarounds

A Misuse Case describes something that is not supposed to happen, a negative scenario which by it's threat can impose new requirements. these scenarios can appear in our case as follows:

- The bee agent can be kidnapped and manipulated by the remote host causing the search result to be spammed with irrelevant results: this scenario is probable since the organizational border of the requester lies directly outside his host, so when an agent migrates to the provider's host, it migrates out of requester's control. However such a misbehavior by the provider can be easily detected using a simple periodical double-checking of returned results. The set of returned results is enough for this checking and the full set of provided services is not needed. When a spamming peer is detected, it can be tracked and the system will learn from the bad experience. Sharing such experience among peers ultimately builds a shared reputation system.
- Remote hosts might not trust the agent's code: this scenario is probable due to the risk of running a foreign code on providers' hosts. However this can be worked around by allowing mobile agents to run within a sandbox with limited permissions; that is the agent framework.
- Communicating the code consumes extra bandwidth: sending the semantic goal as in the traditional way is more bandwidth efficient than sending it together with the code of the agent, and that puts us in front of three modes of communicating the code:
  1. Don't send any code: this mode requires the case that all matching algorithms are implemented by the remote hosts, and the requester has only to choose which one to use. In this mode agents run locally on the requester's host and migrate only virtually; they communicate with providers through conventional network interfaces. Even though this mode can solve the trust issue mentioned above, it features minimal customization possibilities for the requester, and requires all provider peers to implement all matching algorithms.
  2. Always send the code: although this mode features maximal customization possibilities, and releases provider hosts from implementing matching algorithms, it also means unnecessary bandwidth consumption when sending the same agent again and again to the same host.
  3. Send the code only if it isn't cached by the host: This is a hybrid mode, features agent code caching by providers' hosts. Before the agent is sent, the requester inquires about the agent being cached, in the case of yes, it sends only the goal; in the case of no, it sends the goal and the agent's code. This mode has the same customization possibilities as the second, and requires no unnecessary bandwidth consumption.

  These three modes can be used interchangeably in communicating with different peers. However we are arguing that providers will tend to accept the usage of the third mode in order to have a competitive edge.

6 Related Work

The approach discussed in this paper belongs under the research area of adaptive applications. Years of research in this area created two intersecting research communities: ”User Modeling” and ”Adaptive Hypermedia”. Whereas most of the work in these two communities regarded the Web in its current state, some contributions considered the Semantic Web as the upcoming medium for information storage, retrieval and reasoning. In our approach, a simulation of a bee hive is used instead of the more popular simulation of ant colonies, and that was based on works referenced in [2] [3] [4] [5] and [6]. The decision to employ the bee's community simulation rather than ants' was built on the results found by Nyree Lemmens in [13].

7 Conclusion and Future Trends

In this paper we have presented a new beehive-based mechanism for Semantic Web Services discovery in P2P networks as an extension of previous work by the authors [1]; this mechanism is used to personalize the search space in order to get faster and more relevant results. The approach has benefits on four main axes in enhancing the quality of electronic services, further studies are under way to implement a feasibility proof. Although practical results of applying the base model were actually achieved by [4] and more by [5], the extended model is yet to be applied and tested within its environment. An inter-peer social system can incorporate ratings provided by other peers in the evaluation process of certain services; i.e. new rating parameters will be involved as inputs for the fuzzy aggregation model which can in turn be automatically enhanced by introducing a stationary neuro-fuzzy agent that can learn from user feedback.

We also demonstrated the implication and potential benefits of the work in the business domain, how it will improve user interaction with the system, and the level of perceived quality. This can be compared to other currently used systems to show any deviances in the implications of applying a beehive like approach, to solve existing personalization problems which are faced daily by users of web services. Furthermore, we discussed how to utilize the use of such techniques to enhance the search methodology by a manner that will present results based on the users quality metrics and his/her interactivity with the system.

References

1. Memari, A., Brehm, N., Marx-Gomez, J., Mahmoud, T.: Towards intelligent discovery of enterprise architecture services -IDEAS-. Journal of Enterprise Architecture 4(3) (August 2008)
2. Seeley, T.D.: The Wisdom of the Hive: The Social Physiology of Honey Bee Colonies. Harvard University Press (February 1996)
3. Surowiecki, J.: The Wisdom of Crowds. Anchor (August 2005)

4. Navrat, P., Kovacik, M.: Web search engine as a bee hive. In: Proceedings of the 2006 IEEE/WIC/ACM International Conference on Web Intelligence, pp. 694–701. IEEE Computer Society, Los Alamitos (2006)
5. Navrat, P., Kovacik, M., Ezzeddine, A.B., Rozinajova, V.: Web search engine working as a bee hive. Web Intelli. and Agent Sys. 6(4), 441–452 (2008)
6. Lorenzi, F., dos Santos, D.S., Bazzan, A.L.C.: Negotiation for task allocation among agents in case-base recommender systems: a swarm-intelligence approach. In: Proceedings of the Workshop Multi-Agent Information Retrieval and Recommender Systems (July 2005)
7. Tsetsos, V., Anagnostopoulos, C., Hadjiefthymiades, S.: Semantic web service discovery: Methods, algorithms, and tools. In: Cardoso, J. (ed.) Semantic Web Services: Theory, Tools and Applications. IGI Global (2007)
8. Klusch, M., Fries, B., Sycara, K.: Automated semantic web service discovery with OWLS-MX, Hakodate, Japan, pp. 915–922. ACM, New York (2006)
9. Trastour, D., Bartolini, C., Gonzalez-castillo, J.: A semantic web approach to service description for matchmaking of services. In: Proceedings of the International Semantic Web Working Symposium (SWWS), vol. 1 (2001)
10. Fassnacht, M., Koese, I.: Quality of electronic services: Conceptualizing and testing a hierarchical model. Journal of Service Research 9(1), 19–37 (2006)
11. Amer, M., Marx-Gomez, J.: Measuring b2c quality of electronic service: Towards a common consensus. In: Encyclopedia of E-Business Development and Management in the Global Economy (to be published, 2010)
12. Memari, A., Marx-Gomez, J.: A model for adaptive applications on the semantic web. In: 3rd International Conference on Information and Communication Technologies: From Theory to Applications, ICTTA 2008 (2008)
13. Lemmens, N.: To Bee or not to Bee: a Comparative Study in Swarm Intelligence. Master, Maastricht University (2006)

Phonetic Experiment Web Application

Anastazija Kirkova-Naskova[2], Goran Bakraceski[1], Vladimir Apostolski[1],
and Dimitar Trajanov[1]

[1] Faculty of Electrical Engineering and Information Technologies, Karpoš II bb, 1000 Skopje
goran.bakraceski@feit.ukim.edu.mk,
vladimir.apostolski@feit.ukim.edu.mk, mite@feit.ukim.edu.mk
[2] Faculty of Philology "Blaže Koneski" Bul. Krste Misirkov bb, 1000 Skopje
akirkova@t-home.mk

Abstract. This paper describes the development of a methodology and software for phonetics designed with the support of the Internet and web technologies. A web application was created as a data gathering instrument for a phonetic study which aimed to detect the most frequent segmental markers of Macedonian-English accented speech as perceived by native speakers of English and to find out whether English native speakers of different backgrounds perceive the same segments as non-native. The results demonstrate the manifold advantages of the approach as well as the flexibility of its adaptation in applied linguistic research and second language learning/teaching.

Keywords: web application, online experiment, phonetics, computer-assisted research.

1 Introduction

In recent years the incorporation of computer technology in various fields of scientific research has increased enormously and become necessity in itself. One such field where computer assistance marks advancement and yields remarkable outcomes is theoretical and applied linguistics, in particular phonetics, phonology and second language phonological acquisition [4]. Studies related to speech recognition, text-to-speech synthesis, spoken dialogue systems, and speech corpora (recording, archiving, tagging, transcribing) make use of specialist software packages. Furthermore, technology is more and more used to manipulate second-language phonological components in studies related to L2 speech perception and production, foreign accent phenomena, spoken comprehensibility and intelligibility, development of listening and speaking competence as well as to teach particular aspects of L2 pronunciation.

Research practice with foreign accent ratings and segmental analysis studies has shown that traditional methodology of conducting such experiments can be lengthy and wearisome both for the researchers and the participants [2]. For the researchers, this usually involves hours and hours of speech materials preparation. The listeners, on the other hand, are expected to listen to recorded speech stimuli and complete various identification /discrimination tasks or to provide detailed phonetic analysis. As a result, they may become impatient, bored and indifferent thus compromising the validity of the responses. To avoid such problems, researchers started to use computer technology to manipulate speech in order to gain empirical evidence of comprehensibility, fluency and native and non-native perceptions of foreign accents [8, 9]. However, such attempts were made in highly controlled laboratory conditions, with expensive equipment and with limited number of participants.

Only recently have the possibilities that the Internet offers been challenged; apart from general linguistic questionnaires and surveys administered via the Internet and quiz-based applications for practicing phonetic symbols and target sounds, the study conducted under the auspices of the Utrecht Institute of Linguistics OTS in the Netherlands [11] employed a large-scale Internet survey which addressed the issue of intelligibility and acceptability of non-native pronunciation features by English native speakers. In this study the WWStim [13] software is used. WWStim is a perl CGI script for presenting web-based questionnaires and experiments. The system basically presents predefined sequences of template based HTML pages. One of the shortcomings for which we could not use this software was its inability to display phonetic symbols.

In addition to the expensive equipment and the length of the experiments, another drawback with the phonetic studies that collect responses for segmental elements and global accent ratings is the specific listeners' profile required for the experiment and their availability at the time of the study. Flege [5], for example, argues that despite the possibilities of gathering quantitative measurements of foreign-accented speech, it is the qualitative judgments of native speakers that remain "the golden standard". Furthermore, many studies have concluded that ratings gathered from phonetically trained and experienced listeners have proved more reliable [1, 3, 10].

The web application we created was part of a phonetic experiment which was designed to determine and describe the vocalic and consonantal markers of foreign accent in the speech of Macedonian speakers learning English as a foreign language as perceived by English native speakers who speak different English dialects. The choice for such approach was determined by the particular research context with respect to: a) the applicability of traditional methods employed in similar research studies; b) our knowledge of more recent attempts of Internet survey use; and c) the prerequisite for a specific target group of native speakers.

In summary, for the purposes of the phonetic experiment we needed English native speakers speaking different English varieties with linguistic or phonetic expertise; to our misfortune, no such high profile specialists resided in Macedonia. We also wished to address a wider audience and gather relevant linguistic data in a short period. To avoid these shortcomings, the experiment was designed as a web application and eventually proved to be an invaluable tool for gathering authentic listeners' data. We also wanted to design a general purpose software that can be easily adapted for the requirements of future research studies involving language phenomena and in the language teaching/learning practice not only for Macedonian and English but for any other language.

2 Software Solution Description

2.1 Software Requirements

From a technical point of view, we were required to develop a web application that would be able to display a predefined set of questions and keep track of the checked answers by each of its users (namely the English native listeners). In addition, our team had to provide an easy way to play audio files inside the web page itself. Also, the experiment administrator was able to create accounts for the users. Users could choose answers from check box lists or radio button lists, and free text boxes were also available so that the listeners could make comments on speech phenomena they had heard or noticed. The users were not allowed to go backwards with the experiment. They were, nevertheless, allowed to opt for a break during the experiment. To finish the experiment, they were instructed to log in again and continue where they had stopped i.e. by being redirected to the next question. At the end of the experiment, the users were given an option to choose whether they wished to receive an e-mail notification about the results of the phonetic experiment.

2.2 Development Tools and Environments

For the development of the application, we used the Microsoft ASP.NET 3.5 technology, which is especially suitable for developing dynamic web applications, such as the Phonetic Experiment Web Application. The code was written in C# 3.0 and Microsoft Visual Studio 2008 was the development environment which was used in order to design and program the application. The collected data were kept in a Microsoft SQL Server 2005 database. Additionally, a Flash component was developed in order to play the audio files.

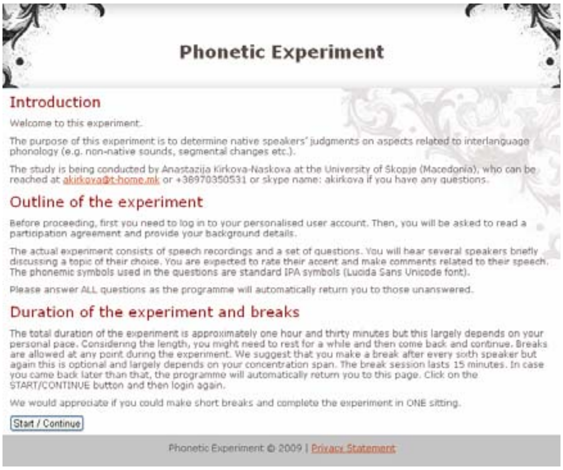
The application work with the most popular Internet browsers available today: Internet Explorer 7+, Mozilla Firefox 2+, Opera or Google Chrome. Because the component that was used to play the audio files was based on Adobe Flash, note that the latest Flash Player was also required.

2.3 Software Design

The idea behind the software solution was to create a general-purpose, flexible poll-based application, adapted for the requirements of the experiment. This way, the application was not dependant on the set of questions that were provided, allowing us to change the number of questions as desired, without changing a single line of code. The same pattern can be reused in various similar applications, not necessarily related to phonetics or, if phonetic in nature, not necessarily for Macedonian and English language. By having the context separated from the logic, we could readapt the same code in different situations with minor changes.

Our data tier was represented by the SQL Server database that we used, the business logic tier consisted of C# classes, which acted as wrappers for the business logic; they were responsible for calculations, binding the questions and the answers and invoking the database queries that were previously written. The presentation tier was represented by the user controls and pages that displayed the results of the business logic that stayed behind them. It is important to note that only one .aspx page was created for the presentation tier, and no query string parameters were passed during the experiment. Session variables were used instead, mainly for security reasons. The session lasted 20 minutes. The audio player Flash component could be observed as add-on to the presentation tier, since it played the mp3 files that were located on the server.

Fig. 1 present screenshots of the phonetics experiment start web pages.



Phonetic Experiment

Introduction

Welcome to this experiment.

The purpose of this experiment is to determine native speakers' judgments on aspects related to interlanguage phonology (e.g. non-native sounds, segmental changes etc.).

The study is being conducted by Anastazija Kirkova-Naskova at the University of Skopje (Macedonia), who can be reached at akirkova@t-home.mk or +38970350531 or skype name: akirkova if you have any questions.

Outline of the experiment

Before proceeding, first you need to log in to your personalised user account. Then, you will be asked to read a participation agreement and provide your background details.

The actual experiment consists of speech recordings and a set of questions. You will hear several speakers briefly discussing a topic of their choice. You are expected to rate their accent and make comments related to their speech. The phonemic symbols used in the questions are standard IPA symbols (Lucida Sans Unicode font).

Please answer ALL questions as the programme will automatically return you to those unanswered.

Duration of the experiment and breaks

The total duration of the experiment is approximately one hour and thirty minutes but this largely depends on your personal pace. Considering the length, you might need to rest for a while and then come back and continue. Breaks are allowed at any point during the experiment. We suggest that you make a break after every sixth speaker but again this is optional and largely depends on your concentration span. The break session lasts 15 minutes. In case you came back later than that, the programme will automatically return you to this page. Click on the START/CONTINUE button and then login again.

We would appreciate if you could make short breaks and complete the experiment in ONE sitting.

Fig. 1. Phonetic experiment start page

2.4 Experiment Flow

A total of 17 Macedonian speakers were recorded producing a free speech in English. They were all Skopje residents, 19-25 years old. All of them were students majoring in English Language and Literature. Their English language proficiency level was B2 (n=6) and C1 (n=11). None of them had a longer stay in an English-speaking country.

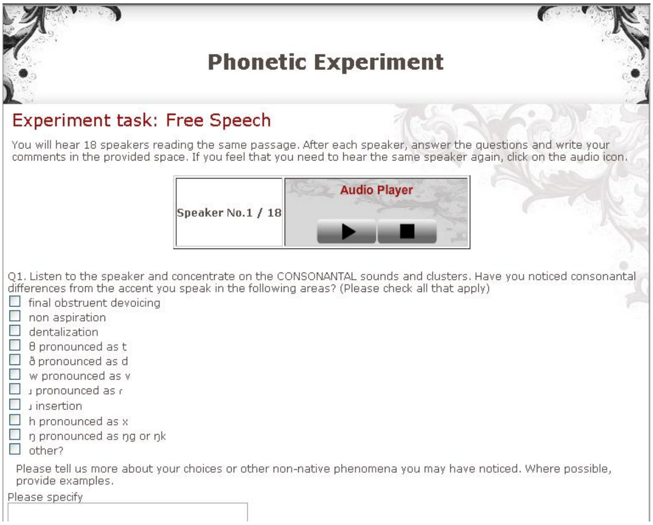
One English native speaker who worked at the Department of English Language and Literature was recorded speaking and was part of the experiment; he served as a control speaker in the accent rating section.

The Macedonian speakers and the control speaker were recorded in a sound prove booth. They were given a 'free speech' task with four optional topics and were expected to speak on the chosen topic for about 2 minutes. The recordings were then edited and tested for authenticity. The total duration of the speech samples that were part of the experiment was 373 seconds (approximately 6 minutes).

The application was advertised as a research project on the LINGUISTLIST http://www.linguistlist.org/ and the IATEFL Pronunciation Special Interest Group - PronSIG yahoo group http://uk.groups.yahoo.com/group/iatefl_pronsig/. These sites are well-known and acknowledged as useful discussion forums as well as media for notification and exchange of information related to current research and conference details. Another method for listener recruitment included direct approach to experts in the field by e-mail and personal acquaintance.

A total of 14 English native speakers completed the experiment and rated the speech produced by the Macedonians. They were 28-71 years old (median 49 years). All of them had gained higher level university degrees and had the phonetic-phonological expertise (some had experience as raters too). As reported in the questionnaire, they spoke the following English variants: Southern British English (n=6), American English (n=6), Irish English (n=1) and Canadian English (n=1).

The web application consisted of several parts organized in separate webpages: 1) introduction and instructions; 2) user account window; 3) participation agreement; 4) personal background details; 5) experiment questions (Q1-Q4) with audio files (repeated for every speaker): consonantal variables, vocalic variables, foreign accent ratings, general variables for foreign accent evaluations; 6) impressionistic comment on Macedonian-English speech; 7) comments on the experiment.



Phonetic Experiment

Experiment task: Free Speech

You will hear 18 speakers reading the same passage. After each speaker, answer the questions and write your comments in the provided space. If you feel that you need to hear the same speaker again, click on the audio icon.

Q1. Listen to the speaker and concentrate on the CONSONANTAL sounds and clusters. Have you noticed consonantal differences from the accent you speak in the following areas? (Please check all that apply)

□ final obstruent devoicing
□ non aspiration
□ dentalization
□ θ pronounced as t
□ ð pronounced as d
□ w pronounced as v
□ ɹ pronounced as ɾ
□ ɹ insertion
□ h pronounced as x
□ ŋ pronounced as ŋg or ŋk
□ other?

Please tell us more about your choices or other non-native phenomena you may have noticed. Where possible, provide examples.

Please specify

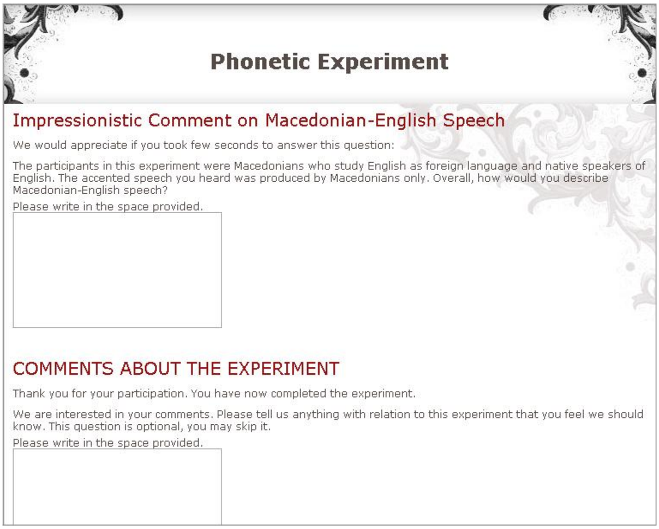
Fig. 2. Question web page

The listeners were expected to listen to the individual speaker's audio file and answer the questions as instructed. The procedure was repeated for every speaker. Two types of data were collected: a) quantitative (frequency of phonetic segment variables and global foreign accent ratings on a 5-point scale), and b) qualitative (open-ended questions with comment boxes).

The administrator first created accounts for each listener and then sent the username and password to the listener via e-mail. Having received the username and password, the listener was expected to log into the system. First, the user had to agree with the terms of usage for the Phonetic Experiment Web Application. Then the application displayed a page with questions rendering data about the listeners' personal data, educational background and prior experience with accent ratings. Once this had been fulfilled, the user was redirected to another page and could start answering the experiment questions. The questions were divided in three parts. An audio file was associated with the set of questions. A list of check boxes and radio buttons was provided for each answer, in addition to a free text box for comments (Fig 2).

The listener answered each question on the page and then had to click on the button to be redirected to the next set of questions (for another speaker). Once the user listened to all speakers and answered the questions, he/she was redirected to a Impressions and comments page (Fig 3). If the listener did not complete the whole experiment in one sitting, he/she could log in again and continue where he/she had stopped, since the application kept track of the user's answers.

The total duration of the experiment was 1-1,5 hours (as reported by listeners who completed the pilot version of the experiment).


Phonetic Experiment

Impressionistic Comment on Macedonian-English Speech

We would appreciate if you took few seconds to answer this question:

The participants in this experiment were Macedonians who study English as foreign language and native speakers of English. The accented speech you heard was produced by Macedonians only. Overall, how would you describe Macedonian-English speech?

Please write in the space provided.

COMMENTS ABOUT THE EXPERIMENT

Thank you for your participation. You have now completed the experiment.

We are interested in your comments. Please tell us anything with relation to this experiment that you feel we should know. This question is optional, you may skip it.

Please write in the space provided.

Fig. 3. Impressions and comments page

2.5 Security Issues

Security is one of the most important aspects when developing web applications. User data must be secure from any threats. For the sake of increasing the level of security, we decided to keep the query string clean and pass no variables through it. That way we ensured that the user could not manipulate the question he/she was answering. Moreover, if the user clicked on the Back button of the browser, he/she could not return to that question because the session variable that was used for tracking the current question was kept on the server, and not on the client. Every time the user clicked on the Submit button, the database was updated as well, meaning the user could leave the application any time he/she wanted.

As every application that collects sensitive data from its users, we published a privacy policy, stating that all data had been used for the phonetic experiment only. It was also stated that cookies had been used, which contained the session keys and other useful data. Our application displayed the privacy policy at the bottom of every page. By using this application, the user agreed to participate on voluntary basis and no payment compensation was given to any participant.

2.6 Data Collection and Data Extraction

One of the reasons for the development of this application was the creation of a small-scale corpus of listeners' responses on segmental aspects related to Macedonian-English speech. This allowed for access to data that would be already stored and eventually easily grouped for faster analysis.

The database behind the application was carefully designed for easy data extraction, meaning tables were connected with each other, containing the primary keys from other related tables, so that SQL queries were much simpler. Once the phonetic experiment was over and all participants had completed it, the data was required to be extracted in a human-readable format. However, the statistical analysis was conducted using SPSS program and this required additional data input to calculate the frequency of checked variables.

3 Results

The satisfactory number of listeners who responded to the project advertisement and their successful completion of the experiment suggest that the Phonetic Experiment Web Application demonstrates a great potential that is yet to be enhanced.

From a linguistic point of view, the experiment in the form of a web application fully met our expectations. The variables used in the questions were typical mispronunciations observed in the phonetic literature related to Macedonian-English speech and in the teaching practice in Macedonia. The result analysis shed light on the predicted phenomena and pointed out to three variables as the most frequent markers of Macedonian-English accented speech: obstruent devoicing in final word position, vowel shortening and substitution of English dental fricatives with Macedonian dental plosives. It also reflected phonetic aspects that we were not aware of or phenomena that used to be poorly explained in the reference literature such as allophonic distributional differences between the two languages and intonational mismatch (sporadic use of weak forms and frequent inappropriate use of rising tones). Based on the listeners' responses (both qualitative and quantitative) we were able to construct a detailed profile of the English speech produced by the Macedonian learner of English (the typical representative of our sample) and to propose practical pedagogical implications.

From a technical point of view, the advantages of the application, being the first of this kind administered in Macedonia, surpass its limitations. As part of the application, the listeners had an opportunity to provide comments about the experiment itself. We have gathered really constructive responses. The positive comments highlighted the clarity of the instructions, the high quality of the recordings in the audio files and the well-defined and user-friendly experiment as a whole. The negative comments addressed the length of the experiment and the need for a pause button in the audio file icon.

The use of the Internet as medium also proved rewarding by making the application global and reachable from anywhere in the world. The targeted group had no problems finding the experiment application. It addressed wider audience i.e. we had respondents from the USA, Canada, the UK, Ireland, New Zealand . It proved to be time-saving rather than time-consuming both for the researcher and the users/listeners. There was no need for the researcher to directly supervise experiment completion as the procedure was pre-programmed and could be administered on various locations at the same time, and the users were allowed to conduct the experiment according to their own schedule. Another gain was the speed and efficiency of data collection, the researchers receiving immediate results and a database being instantly created and regularly updated. This also meant that data was easily managed because there was no need for manual data input. Most importantly, the application can be easily adapted for the requirements of similar research studies due to its flexible software design.

4 Conclusion

Phonetic Experiment Web Application is an example of how a modern approach can be applied when conducting linguistic experiments. There is room for improvement and adaptation not only for research purposes but also in the area of language teaching, learning and assessment.

Our team hopes that this application will become popular with online linguistic communities and improve the methodology employed in linguistically-related experiments. In the educational context, this application can be modified as part of learning management software where teacher-student interaction is preferred. The students can record themselves and upload their speech as an audio file. The teacher on his/her part can mark their mispronunciations, give immediate feedback, and monitor their progress. Alternatively, as part of their exams, students may be required to record themselves and upload the file and the teacher can assess their pronunciation on a set of predefined phonetic items that are expected to be acquired throughout the academic year.

Finally, this approach is ideal for the promotion of research in lesser-developed countries, as is the case with Macedonia, where people and resources are always limited and insufficient. This positive experience also shows the continuing need for theoretical and applied linguists to collaborate with computer technology experts in order to improve methods and bridge the gap between research and practice.

References

[1] Anderson-Hsieh, J., Koehler, K.: The effect of foreign accent and speaking rate on native speaker comprehension. Language Learning 38(4), 561–613 (1988)

[2] Beddor, P.S., Gottfried, T.S.: Methodological issues in cross-language speech perception research with adults. In: Strange, W. (ed.) Speech Perception and Linguistic Experience: Theoretical and Methodological Issues in Cross-Language Speech Research, pp. 207–232. York Press, Baltimore (1995)

[3] Bongaerts, T., van Summeren, C., Planken, B., Schils, E.: Age and ultimate attainment in the pronunciation of a foreign language. Studies in Second Language Acquisition 19, 447–465 (1997)

[4] Chun, D.M.: Technological advances in researching and teaching phonology. In: Pennington, M.C. (ed.) Phonology in Context, pp. 274–299. Palgrave Macmillan, Basingstoke (2007)

[5] Flege, J.E.: Factors affecting the pronunciation of a second language [.ppt, electronic version]. Presented at Pronunciation Modelling and Lexicon Adaptation for Spoken Language Technology, Estes Park, Colorado, USA, September 14-15 (2002), http://jimflege.com/files/Colorado_2002.pdf

[6] Flege, J.E., Munro, M.H., MacKay, I.R.A.: Factors affecting strength of perceived foreign accent in a second language. Journal of the Acoustical Society of America 97, 3125–3134 (1995)

[7] Major, R.C.: Paragoge and degree of foreign accent in Brazilian English. Second Language Research 2, 53–71 (1986)

[8] Munro, M.J., Derwing, T.M.: Foreign Accent, Comprehensibility and Intelligibility in the Speech of Second Language Learners. In: Leather, J. (ed.) Phonological Issues in Language Learning, pp. 285–310. Basil Blackwell, Oxford (1999)

[9] Munro, M.J., Derwing, T.M.: Modelling perceptions of the Accentedness and Comprehensibility of L2 Speech: The role of speaking rate. Studies in Second Language Acquisition 23, 451–468 (2001)

[10] Thompson, I.: Foreign accents revisited: The English pronunciation of Russian immigrants. Language Learning 41(2), 177–204 (1991)

[11] Van den Doel, R.: How friendly are the natives? An evaluation of native-speaker judgments of foreign-accented British and American English. LOT, Utrecht (2006)

[12] Weinberger, S.H.: The speech accent archive [electronic database]. George Mason University (1998), http://accent.gmu.edu/index.php

[13] WWStim, http://www.let.uu.nl/~Theo.Veenker/personal/projects/wwstim/doc/en/

Using XAML in Representation of Dental Charts in Electronic Health Record

Ivica Marković[1], Srebrenko Pešić[2], and Dragan Janković[1]

[1] Faculty of Electronic Engineering, University of Niš, Aleksandra Medvedeva 14,
18000 Niš, Serbia
{ivica.markovic,dragan.jankovic}@elfak.ni.ac.rs
[2] Health Care Center Niš, Vojvode Tankosića 15, 18000 Niš, Serbia
srebrenko.pesic@domzdravljanis.co.rs

Abstract. Current technology used for displaying Windows forms is about 15 years old and it is based on two parts of the Windows operating system - User32 library and GDI/GDI+ API. Microsoft's new technologies in this area, WPF and XAML, improve current situation by offering better optimization of code execution, extended code reusability and a fresh new visual appearance. New technologies improve quality of created GUI and increase the usage area. Usage of these new technologies in implementing GUI for dental charts application which is developed as a part of a broader medical information system is described in this paper. Achieved results confirm improved quality which is brought by new technologies.

Keywords: User Interface Design, WPF, XAML, Dental Charting.

1 Introduction

From the very beginning of programming languages development primary goal was to make programming more efficient. During all this years, programming languages were constantly improved. The ideal language should provide possibility to describe solutions naturally and hide unnecessary details. Also, it should be expressive enough in the problem domain and should provide guarantees on properties critical for the problem domain. Developing of domain specific languages was always good solution for specific problems which couldn't be solved by using general-purpose languages. Specific problem which is addressed in this paper is development of graphical user interfaces (GUI).

In last decade, Extensible Markup Language (XML) takes part in data exchange between different applications, file formats etc. XML is hard for humans to read and write, but it seems that is very well accepted by leading technologies in software industry. This acceptance is visible in almost all technologies by supporting XML formats in different use cases. XML usage and its tool support are spreading. One of new aspects is using XML in representing graphical user interfaces. Extensible Application Markup Language (XAML), is a language for representation of GUI in Windows Presentation Foundation (WPF) and Silverlight applications of .NET 3.5, as described in [6] and [7].

It is very important to mention that technology used for displaying Windows forms is 15 years old [1]. A standard Windows application relies on two well-worn parts of the Windows operating system to create its user interface:

- User32 provides the familiar Windows look and feel for elements such as windows, buttons, text boxes, etc.
- GDI/GDI+ provides drawing support for rendering shapes, text, and images at the cost of additional complexity.

Over the years, both technologies have been refined, and the application programming interfaces (APIs) that developers use to interact with them have changed dramatically. But whether you are crafting an application with .NET and Windows Forms, or lingering in the past with Visual Basic 6 or MFC-based C++ code, behind the scenes the same parts of the Windows operating system are at work. Newer frameworks simply deliver better wrappers for interacting with User32 and GDI/GDI+. They can provide improvements in efficiency, reduce complexity, and add new features so you don't have to code them yourself; but they can't remove the fundamental limitations of a system component that was designed more than a decade ago.

Microsoft created one way around the limitations of the User32 and GDI/GDI+ libraries: DirectX. DirectX began as toolkit for creating games on the Windows platform. Its design mandate was speed, and so Microsoft worked closely with video card vendors to give DirectX the hardware acceleration needed for complex textures, special effects such as partial transparency, and three-dimensional graphics.

Medical information systems represent a specific group of information systems with requirements for carefully designed user interface. Dental chart application is one of more demanding parts of medical information systems regarding GUI. Beside standard GUI components for data manipulation (such as buttons, text boxes, combo boxes, data grids) an efficient, interactive tool for graphical modeling of patient's teeth state is needed.

In this paper we present our implementation of Windows forms drawing when advanced controls and animations are desired for end-user convenience. The basic idea of our work was to avoid these difficulties by using XAML and WPF. Section two describes basics of new concepts introduced in WPF related to GUI. In section three one part of large medical information system which uses concepts from section two is described. It is shown there that usage of new technologies has led to GUI implementation which satisfies end-user needs. The last part contains conclusions and plans for future improvements.

2 XAML Basics

The only thing that changed over the past 10 years in Windows display system is DirectX component. But, because of its raw complexity, DirectX is almost never used in traditional types of Windows applications.

Windows Presentation Foundation changes all this. In WPF, the underlying graphics technology is not GDI/GDI+. Instead, it is DirectX. Remarkably, WPF applications use DirectX no matter what type of user interface it is created. That means that whether programmer is designing complex three-dimensional graphics or just drawing buttons and plain text, all the drawing work travels through the DirectX pipeline. As a result, even the most simple business applications can use rich effects such as transparency and anti-aliasing. There is also benefit from hardware acceleration, which simply means DirectX hands off as much work as possible to the graphics processing unit (GPU), which is the dedicated processor on the video card.

Extensible Application Markup Language is a markup language used to instantiate .NET objects. Although XAML is a technology that can be applied to many different problem domains, its primary role is to construct WPF user interfaces. In other words, XAML documents define the arrangement of panels, buttons, and controls that make up the windows in a WPF application.

With traditional display technologies, there is no easy way to separate the graphical content from the code. The key problem with Windows Forms application is that every form is defined entirely in C# code. As programmer drop controls onto the design surface and configure them, Visual Studio quietly adjusts the code in the corresponding form class. Unfortunately, graphic designers do not have any tools that can work with C# code. WPF solves this problem with XAML. When designing a WPF application in Visual Studio, the window that is designed is not translated into code. Instead, it is serialized into a set of XAML tags. When the application is run, these tags are used to generate the objects that compose the user interface.

It is unlikely that human will write XAML by hand. Instead, they will use a tool that generates the XAML that they need. If they are graphic designers, that tool is likely to be a graphical design and drawing program. If they are programmers they will probably use Visual Studio.

It is important to mention that WPF does not require XAML. There is no reason Visual Studio could not use the Windows Forms approach and create code statements that construct WPF windows. But if it did, that window would be locked into the Visual Studio environment and available to programmers only.

The creators of WPF knew that XAML needed to not just solve the problem of design collaboration, but also needed to be fast. And though XML-based formats such as XAML are flexible and easily portable to other tools and platforms, they are not always the most efficient option. XML was designed to be logical, readable, and straightforward, but not compact. WPF addresses this with Binary Application Markup Language (BAML). BAML is a binary representation of XAML. When a WPF application is compiled in Visual Studio, all XAML files are converted into BAML and that BAML is then embedded as a resource into the final DLL or EXE assembly. BAML is tokenized, which means lengthier bits of XAML are replaced with shorter tokens. Not only is BAML significantly smaller, it is also optimized in a way that makes it faster to parse at runtime. Most developers will not worry about the conversion of XAML to BAML because the compiler performs it behind the scenes. Also, it is possible to use XAML without compiling it first. This might make sense in scenarios that require some of the user interface to be supplied just-in-time (for example, pulled out of a database as a block of XAML tags).

3 XAML in Medis.Net.Dental Application

Medis.Net.Dental is a part of a medical information system (MIS) which is being developed at The Faculty of Electronic Engineering in Niš (http://medisnet.elfak.ni.ac.rs) in cooperation with The Health Care Center in Niš. Development of medical information system is a central part of a project supported and funded by The Ministry of Science and Technological Development of Republic of Serbia.

Main task of Medis.Net medical information system for ambulatory care facilities (Fig. 1) is direct mapping of all administrative procedures to information system with important improvement in communication between them. For each paper based document a related look-a-like Windows form is created to make the system easier to use for medical personnel, since they will not change their standard procedures, except using MIS instead of filling paper documents. Medis.Net consists of central database and several client modules. There are both modules common for all users (such as modules for authorization and appointments) and user specific modules (such as general purpose record, dental record, specialist's workspace). Authorized users are allowed to access application modules corresponding to their work area and patients' data for their patients only. More details about this medical information system can be found in [6].

One of user specific modules in Medis.Net is Medis.Net.Dental, a software solution for dental clinics and dental practices (Fig. 2). It allows its users to handle the data about their patients' teeth current state, history of the patients' teeth state, applied therapy, used materials, invoices for given services, work schedule of medical personnel, patients' appointments etc.

Very important functionality of Medis.Net.Dental is managing descriptions of patients' teeth state by careful charting of both visual (Fig. 4) and radiographic findings (Fig. 3) which allow dental practitioners to collate information needed to assess the patient's level of dental and periodontal health or disease. For majority of diagnoses and therapies radiographic images are not necessary or even undesirable because in process of their creation patient is exposed to radiation. Therefore central part of this application is patient's dental chart which is mostly filled according to the visual findings of the dentists.

Dental charting provides a graphic description of the conditions in a patient's mouth, including caries (i.e., decay), restorations, missing or malposed teeth, furcation (root) involvement, mobility, pocket depths and other deviations from normal. Other conditions that may be charted include use of prostheses, dental implants, bridges, crowns and other restorative procedures [2]. Charting should be updated with each visit to follow the patient's progress with home care, monitor disease progression, and to track completed dental procedures. In that way a comprehensive history view of patient's teeth state will be available to the dentist during patient's future visits. Teeth history view is available as a preview of all modifications of one tooth or as a chart of all teeth on a specified date.

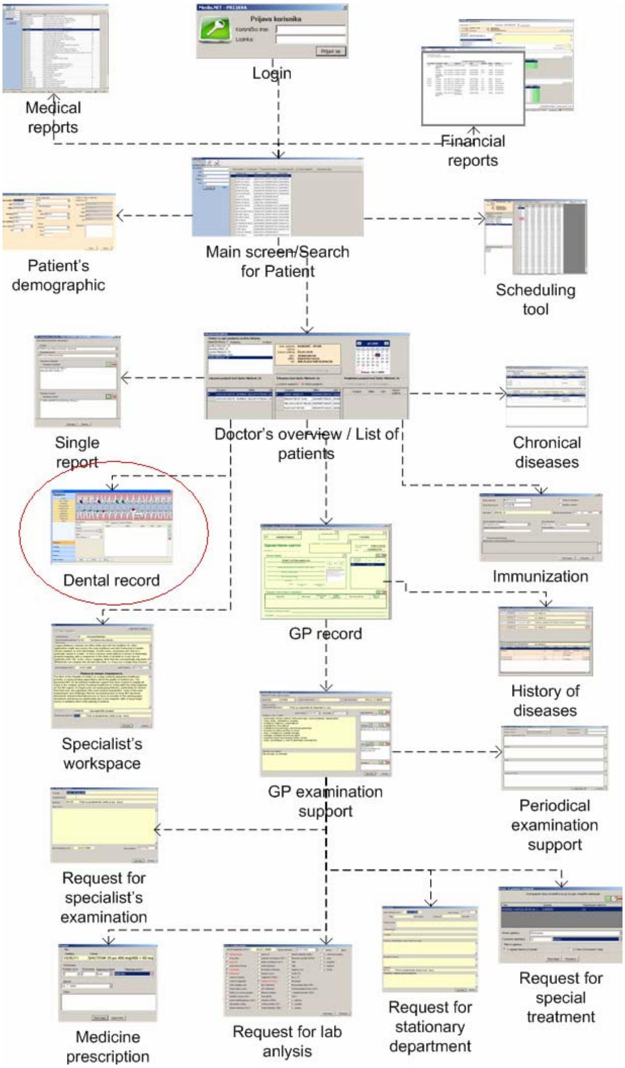
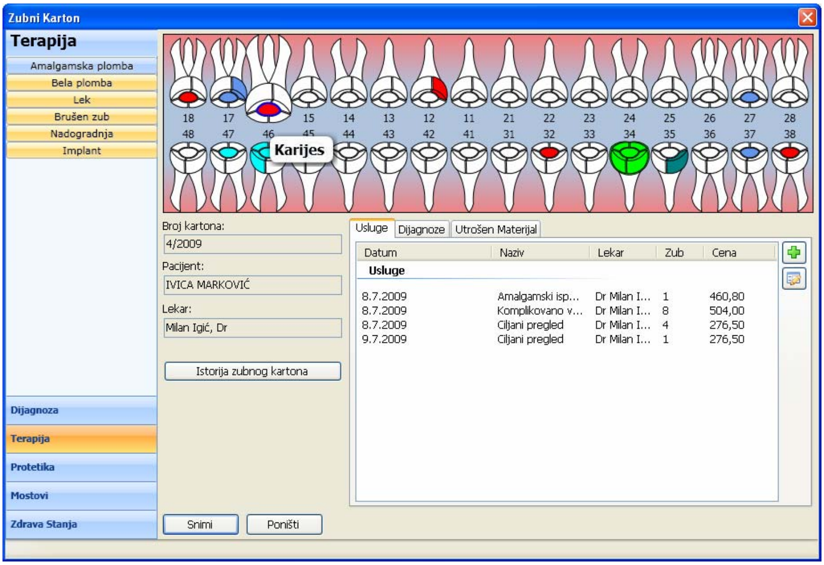


Fig. 1. Diagram showing position of Medis.Net.Dental in complete medical information system

Medis.Net.Dental application (Fig.2.) is implemented as a .NET 3.5 Windows form application in C# programming language. It is implemented by using standard windows forms mechanism from .NET Framework with well known controls. Only for implementation of graphically more complex controls such as tooth chart control in upper part of the form (Fig.4.) and outlook bar in the left WPF and XAML were used. This form can be used as an example to show differences in visual appearance between controls rendered in old and new technology.



Broj kartona: 4/2009

Pacijent: IVICA MARKOVIĆ

Lekar: Milan Igić, Dr

Datum	Naziv	Lekar	Zub	Cena
Usluge				
8.7.2009	Amalgamski isp...	Dr Milan I...	1	460,80
8.7.2009	Komplikovano v...	Dr Milan I...	8	504,00
8.7.2009	Ciljani pregled	Dr Milan I...	4	276,50
9.7.2009	Ciljani pregled	Dr Milan I...	1	276,50

Fig. 2. Basic dentist's view in Medis.Net.Dental application. The most important part of this view is tooth chart control for graphic description of patient's teeth status.

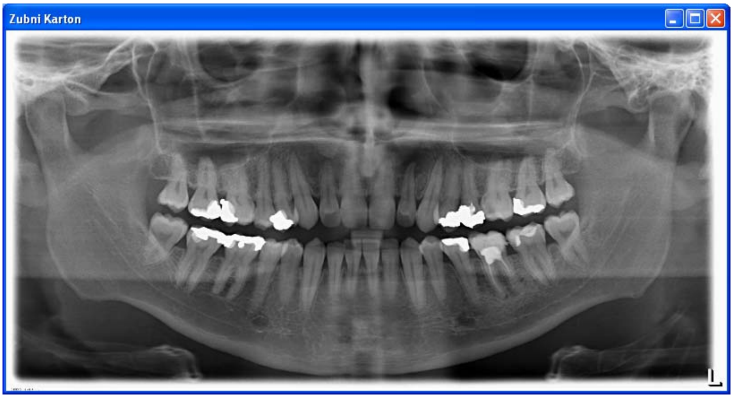


Fig. 3. A view of radiographic image in Medis.Net.Dental application

XAML implementation of tooth chart control (Fig. 4.) was separated into 3 classes:

- ToothPieceShape,
- Tooth,
- Teeth.

ToothPieceShape class is used for drawing single tooth part such as root or crown part (facial, mesial, distal, lingual or incisal surface). It extends Shape class from WPF and mostly exploits methods for drawing graphical primitives from StreamGeometryContext class such as BeginFigure(),LineTo(),ArcTo(), etc. for drawing tooth parts. Equivalent primitives are also available in GDI/GDI+ but WPF methods are in advantage because of performing the drawing by specialized graphics hardware. Using GDI/GDI+ would waste CPU time for these operations, possibly leading to application becoming slow and non-responding.

Tooth class is used for drawing a tooth represented as a collection of ToothPieceShape objects. It is implemented as a combination of both XAML and C# code. XAML code (see the code listing bellow) defines design related features such as color scheme, shadow and opacity effects, layout of roots and crown parts and appearance of tooltips. Methods for zoom in / zoom out tooth, changing states (diagnoses and therapies) of tooth parts and event handlers are written in C# and are using WPF libraries. This enabled nice and smooth animations when in a MouseEnter event a tooth is zoomed in and when in a MouseLeave event a tooth is zoomed out.

Teeth class is used for drawing tooth chart represented as a collection of Tooth objects. It also combines XAML and C# code and, similarly to Tooth class, in XAML part layout of teeth and color scheme were defined.

Outlook bar control was less complex to implement and only one class was used. It is almost completely written in XAML.

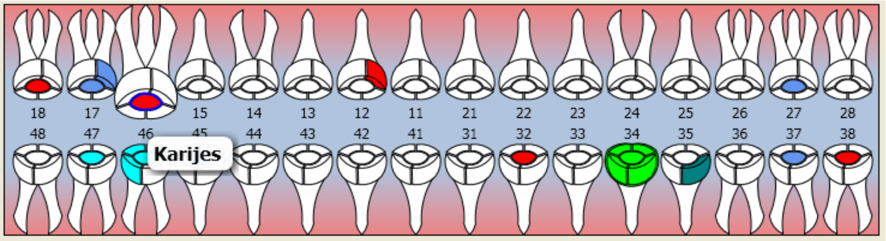


Fig. 4. Tooth chart control for graphic description of patient’s teeth. The control was developed using XAML.

In all classes with XAML code some new features and concepts such as dependency properties and routed events were used.

A XAML code fragment in class Tooth which represents a single tooth

```
<UserControl x:Class="Medis.Dental.Tooth">
 <UserControl.Resources>
  <Style TargetType="{x:Type ToolTip}">
  <Setter Property="Opacity" Value=".95"/>
  <Setter Property="Template">
   <Setter.Value>
    <!-- modify the tooltip control template to
         add a drop shadow-->
    <ControlTemplate TargetType="{x:Type ToolTip}">
     <Grid Background="Transparent" Margin="5"
           Width="{TemplateBinding Width}"
           Height="{TemplateBinding Height}">
      <Rectangle Fill="White"
                 Height="{TemplateBinding Height}"
                 RadiusX="7.5" RadiusY="7.5">
       <Rectangle.BitmapEffect>
        <DropShadowBitmapEffect ShadowDepth="3"/>
       </Rectangle.BitmapEffect>
      </Rectangle>
      <ContentPresenter Margin="5"
                        HorizontalAlignment="Center"
                        VerticalAlignment="Center" />
     </Grid>
    </ControlTemplate>
   </Setter.Value>
  </Setter>
  <Setter Property="ContentTemplate">
   <Setter.Value>
    <DataTemplate>
     <!-- bind the stack panel datacontext to
         the tooltip data context -->
     <StackPanel Orientation="Horizontal"
                 DataContext="{Binding Path=DataContext,
                  RelativeSource={RelativeSource
                  AncestorType={x:Type ToolTip}}}">
      <!-- navigate to the pie piece and obtain
         the percentage -->
      ...
     </StackPanel>
    </DataTemplate>
   </Setter.Value>
  </Setter>
 </UserControl.Resources>
</UserControl>
```

In addition to using regular Common Language Runtime (CLR) properties, WPF also uses dependency properties. Dependency properties use more efficient storage and support higher-level features such as change notification and property value inheritance (the ability to propagate default values down the element tree). Dependency properties are also the basis for a number of key WPF features, including animation, data binding, and styles. A dependency property can only be created in a class which derives from the System.Windows.DependencyObject class. That class is very high up in the WPF class hierarchy, which allows the majority of classes in WPF to use dependency properties. Properties in WPF can be accessed both from C# code and XAML (as it can be seen in the code listing above). In Medis.Net.Dental collection of Tooth class instances is set as a dependency property in Teeth class to trigger re-drawing of Teeth object each time when a Tooth is modified.

The second new feature is higher-level events concept (besides ordinary .NET events) called routed event feature. Routed events are events with more traveling power – they can be transferred down or up the element tree and be processed by event handlers along the way. Routed events allow an event to be handled on one element (such as Teeth class) even though it originates on another (such as ToothPieceShape class).

Described WPF controls are easily integrated into Windows forms application by using ElementHost class from WindowsFormsIntegration assembly. This work is done automatically by Visual Studio designer after dragging a WPF control to a Windows form. Also it can be done manually by instantiating an ElementHost object and setting its property Child to point to the WPF control.

In addition to good interoperability between WPF and .NET Windows forms, it should be mentioned that adopting of new technologies by developers already familiar with old ones didn't significantly slow down development of this application. Some more studies about programmers' experiences with XAML can be found in [3].

4 Conclusion and Future Work

Current technology used for displaying Windows forms, based on User32 and GDI/GDI+, although being widely used and well known among developers has its weakness. Therefore we combined them with Microsoft's new technologies, WPF and XAML, in our work on Medis.Net.Dental application and found several benefits.

Benefits of using XAML in this case are compact and readable source code, fast and smooth animations as responses to user actions without significant CPU time consumption as well as source code reusability. Currently, Medis.Net.Dental is implemented as a desktop client application. In certain cases for a large number of users web application could be in advantage over desktop client application (no need for installation, easy software update). If such architectural modification is required, XAML provides a nice feature that GUI components developed in a Windows application can be completely reused in web applications [4].

Next step in our work should be replacing currently used Windows forms with new components developed in XAML. Expected effect of suggested modification is certain performance improvement and it should be examined by measuring performances and consumption of resources in both applications. This modification will also open up possibility of relatively straightforward transformation of Medis.Net.Dental into a web application in order to maximize number of users. Here presented desktop application forms should preserve their current look although will be rendered in different web browsers.

Acknowledgments. Presented solutions for medical information systems are results of projects supported and funded by the Ministry of Health and the Ministry of Science and Technological Development of the Republic of Serbia (Project Nr TR13015).

References

1. MacDonald, M.: Pro WPF in C# 2008: Windows Presentation Foundation with.NET 3.5, 2nd edn. Apress, Berkeley (2008)
2. Woodall, I.R.: Comprehensive dental hygiene care, CV Mosby, St Louis (1993)
3. Mernik, M., Kosar, T., Crepinsek, M., Rangel Henriques, P., Da Cruz, D.: Comparison of XAML and C# Forms using Cognitive Dimension Framework. Accepted for INFORUM 2009, Lisboa (2009), http://inforum.org.pt/INForum2009/programa
4. Martinez Ruiz, F.J., Muñoz Arteaga, J., Vanderdonckt, J.: Transformation of XAML schema for RIA using XSLT & UsiXML. Proc. of XlX Congreso Nacional y V Congreso Internacional de Informática y Computación de la ANIEI. In: Avances en Tecnologías de la Información CNCIIC 2006, Tuxtla Gutiérrez (2006)
5. Rajković, P., Janković, D., Tošić, V.: A Software Solution for Ambulatory Health Facilities in the Republic of Serbia. Accepted for Healthcom 2009. 11th International Conference on e-Health Networking, Application and Services, Sydney (2009)
6. XAML Overview at MSDN.NET Framework Developer Center, http://msdn.microsoft.com/en-us/library/ms752059.aspx
7. Windows Presentation Foundation at MSDN .NET Framework Developer Center, http://msdn.microsoft.com/en-us/library/ms754130.aspx

Enterprise Tomography Driven Governance of Federated ERP in a Cloud

Jan Aalmink, Lama Balloul, Jan Glagau, and Jorge Marx Gòmez

Business Informatics I, Very Large Business Applications,
Carl von Ossietzky University,
Ammerlaender Heerstr. 114-118
26129 Oldenburg, Germany
{balloul,glagau,marx-gomez}@wi-ol.de, ent.tomo@googlemail.com
http://vlba.wi-ol.de

Abstract. Enterprise Cloud Computing becomes more and more prevalent in the IT and Business Application Industry. The scientific approach is now, to overcome most of the disadvantages of legacy on-premise solutions. Therefore, the existing different research streams, requirements and semantic perspectives need to be converged into one central ubiquitous, standardized architectural approach. The goal is to perform on-demand and cross-enterprise business processes in the context of Very Large Business Applications (VLBAs). Also in this context cloud standardization is one of the biggest challenges of the Open Cloud Manifesto. This paper discusses and outlines, how a semantic composition and federation based reference model (federated ERP-system) can be established for Enterprise Cloud Computing and set up for business operation. Furthermore, it is debated, how enterprises can develop and maintain enterprise software solutions in the Cloud Community in an evolutionary, self-organized way complying to Cloud Standards. In this context a metric driven Semantic Service Discovery and the Enterprise Tomograph can be seen as an entrypoint to an organic, gradable marketplace of processes exposed by cloud based Service Grids and Data Grids in graded levels of granularity and semantic abstractions.

Keywords: Federated ERP, Enterprise Tomography, Cloud Computing, Green Cloud, Enterprise 2.0, Semantic Service-Oriented Architecture.

1 Introduction

Regarding Enterprise Cloud Computing, conflictual requirements and design principles need to be resolved. A convergence of the polymorphic streams towards a shared, cloud-based platform can be observed. The main motivation in utilizing Enterprise Cloud Computing for a customer is the reduction of TCO in different aspects: Pooling of resources and services regarding consumption peaks or simplification of legacy infrastructure from onpremise solutions towards an on-demand solution. From the perspective of an Enterprise Cloud Provider virtualization with multi-tenancy functionality proves as suboptimal. There is a higher degree of sharing and reuse possible. This leads to federated service-based cloud software which can grow organically. The scientific challenge is to provide a controllable reference model which serves as a common standard, where standards overcome the typical vendor-lock-in phenomenon and are prerequisite for acceptance.

In general, Federated ERP Systems (FERP Systems) based on web-services are heterogeneous software systems processing business data complying integration rules, so different customers have different views, i.e. access points to the FERP. Since the typical software ownership (provider-consumer) is transformed from 1:n to m:n [1,2] and the complexity of such information eco-systems is increasing in the course of the life cycle, the superordinate target in the context of Enterprise Cloud Computing is to provide methodologies and mechanisms for streamlining and controlling the integration in federated ERP systems. The organic growth of interlinked Enterprise Services Network needs to follow compliance rules. Therefore semantic deviation-analysis of enterprise service consumption, Monitoring, Tracking becomes essential in distributed consumer-provider networks along the life cycle.

The Enterprise Tomography approach enables the monitoring of the complete life cycle of a federated Enterprise Software. With Enterprise Tomography it is possible to make consumption patterns comparable. This comparison is based on a common interlingua represented as lightweight ontologies and is achieved by applying Delta Operator which determine the differences between system-status A and system-status B in a cloud. To be more precise, the comparison and evolution-tracking of integrated business process scenarios in a cloud represented as interlinked enterprise services ensembles is possible. The Enterprise Tomography approach provides the possibility to visualize differences with help of tomograms which aggregate indicators, metrics and serves as a decision basis in the governance process and Integration LifeCycle Management of an Enterprise Cloud [3].

Figure 1 illustrates an overview of the topology of a Cloud Farm. Different aspects and fundamental pillars of FERP reference model are shown. The procedure, how Enterprise Cloud Evolution can be controlled is outlined in Figure 2.

2 Federated ERP in a Cloud

In reality, on-premise Standard Enterprise Software is widely used within the enterprise community. Standard means, there are common business process patterns which are highly configurable and extensible according to the business requirements. Typically, this methodology results in similar composed, configured and enhanced Enterprise Software Systems deployed to many Enterprise Data Centers. Similarity means redundancy, which can be eliminated with the FERP approach.

According to Figure 2, in the cloud-based Federated ERP approach we have one single software instance active for all participating enterprises. Each enterprise is encapsulated in a Cloud Tenant according to the Separation of Concerns

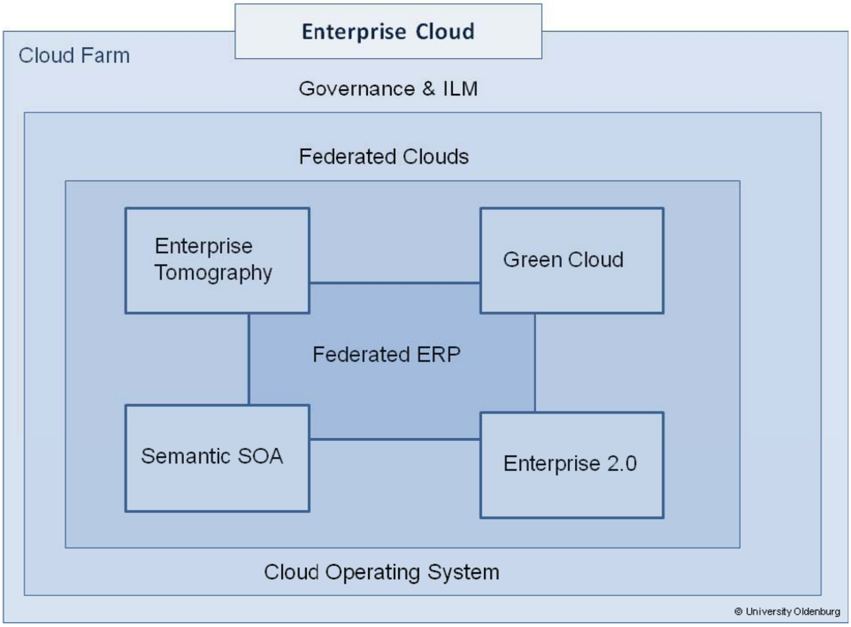


Fig. 1. Topology of an Enterprise Cloud Computing Farm based on FERP and Enterprise Tomography

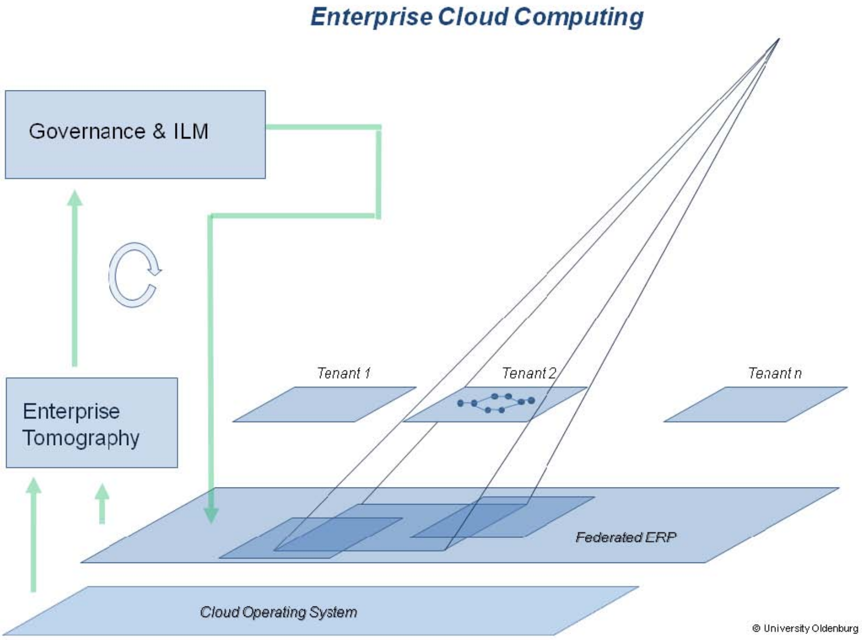


Fig. 2. Enterprise Tomography driven Governance of FERP in a Cloud Farm

Paradigm. Each Tenant is provided a view on the single software and data instance. Basically, the software and data instance is a network of shared Business Objects that are projected on columnar In-Memory databases [6]. The In-Memory Databases can be regarded as intelligent Caches [4]. In-Memory Columnar Databases significantly reduces the redundancy in the data volume and provides instantaneous access to non-materialized aggregates and business object collections. Aggregates are being calculated on the fly and are exposed as services via endpoints of Data Grids. It is possible to keep the Business Object Network consistent according to the ACID transactional OLTP methodology. Columnar In-Memory Database Models provide extensibility by nature.

Non-frequent used Business Objects are physically stored in a distributed fashion. A read access of a Business Object means data retrieval of distinct fragments for reconstruction of the original Business Object. A Business Object is regarded as a tree serialized to a document. This document is fragmented. The fragments are coded and distributed within the Data Grids. A document can be seen as a sequence of numbers which defines a mathematical polynom. According to the fundamental theorem of algebra, this document can be uniquely reconstructed, if there are only n distinct fragments (out of a redundant coded set) available. While retrieving, inconsistent fragments can be ignored and substituted by distinct consistent fragment retrieved from remote Data Grids [5].

In the FERP approach technical references to Business Objects are the payload of messages. E.g. if company A wants to send company B an invoice (Business Object) only the reference of the Invoice is sent as a payload. The invoice is in this case a shared and ubiquitous accessible Business Object. Company A has an individual view-based access to the Invoice via the reference only. The same applies to company B. Receiving the message, company B will change the status of the invoice to the value 'paid' as soon as the real payment is executed.

Columnar Databases are based on Inverted Indexing known in classical Information Retrieval. In [6] it is shown that this algorithmic approach is well-suited for parallel multicore hardware. Systolic Arrays are in the position to accelerate string position/value matching even further with the rate of clock frequency speed [7].

View-based access via references to Business Objects has the big advantage, that no mapping and technical transformation of the Business Object is required. Business Objects needs not to be moved within the memory. There is no need for asynchronous processing and updating anymore. This leads to tremendous scalability which is a prerequisite for cloud computing. Having instantaneous services in places, completely new quasi real-time applications will be possible in future.

In addition, the FERP reference model leads to a more data-consistent behavior. The cloud software can become much more lean in comparison to classical stacked on-premise enterprise software solutions are therefore less error-prone. A closed-loop feedback development process ensures a promptly iterative correction cycle. This leads to quality ensurance.

3 Enterprise Tomography Driven Governance of Integration Lifecycle Management

The Federated ERP model can be regarded as a central shared and ubiquitous accessible network based approach. An error in the enterprise cloud software can lead to dramatic consequences and might have serious business impact.

An Enterprise client can extend its own business processes or even create and compose its individual business process schemas. The individual part of functionality can be shared with related tenants. So FERP leaves the classical Software Vendor / Software Client ownership model. In FERP approach each individual Tenant can be simultaneously in the role of a service consumer as well as a logical service provider. The services are exposed via a Semantic Service Discovery [4]. The essential point here is, that each service, composite service or business process is potentially provided with a set of alternatives distinguished by Quality of Service (QoS) and metrics.

To be more general, the FERP approach can be seen as a definition of a governed service marketplace. Each individual participant can contribute materialized cloud content as shared services and shared (sub)-processes. Each participant can virtually compose his own ERP. In fact he gets a view on an service of an one software and data instance.

With Enterprise Tomography it is possible to make similar data contexts comparable. The comparison is based on a common Interlingua represented as lightweight ontologies. With a Delta Operator it is realizable to determine dynamically the differences between Service offering A and Service offering B. The Enterprise Tomograph provides the possibility to visualize semantic differences with help of tomograms. A comparison between two service offerings is possible as well as a comparison of a service offerings between two points in time. E.g. in a project a consulting team implements business processes and therefore changes Customizing or the alters the composition of a Enterprise Service Ensemble. This delta is of common interest. E.g. as an indicator for the quality of security evolvement in the last period of time.

An other use-case is to determine the delta after an functional upgrade in the cloud. The delta is calculated between the previous reference version and the active version of Enterprise Service Ensemble. The delta in this case is the equivalent of new or changed functionality. This delta, represented as an hierarchical ontology tree, is a good basis for evaluation of new functionality. Test and training teams therefore can focus on new/changed functionality only. This results automatically in cost containment.

One more interesting use case for Enterprise Tomography is to calculate the data footprint of a selected business transaction or a business process in a cloud. Between two points in time the update on database is calculated with help of the Delta Operator. Based on the business data delta, the IT experts are in the position to assess the correctness of the behavior of the executed business transactions more efficiently. This is an highly efficient diagnostics approach for root cause analysis for given error symptoms. Based on the delta, the Undo Operator resets the business transaction. This business transaction can executed again with same preconditions and data contexts. In this way repetitive testing of business processes is enabled.

The Enterprise Tomography approach allows the construction of an early warning system based on semantic metrics and indicators. If the distance - computed by the delta operator - exceeds a threshold, actions (= cloud based services) can be executed to control the usage of dedicated Enterprise Services. For example, the Enterprise Tomograph can execute process mining. When the quota exceeds a threshold, the Enterprise Tenant needs to be invoiced for funding the cloud infrastructure he has used. This is a simple example to implement selforganized feedback control system based on the generic Enterprise Tomography approach.

Each participant can contribute service based software as materialized cloud content. This naturally leads to high redundancy in offerings of business processes. The Enterprise Tomograph can evaluate the services and business processes according real consumption patterns. Business processes with low traffic on the cloud infrastructure are regarded as nonvalue added processes and will be disabled. The decision of disablement is based on dynamic calculated results of the Enterprise Tomograph. The most useful services - or more general - the services with the highest Quality of Services will survive the market competition. This example illustrates, how Enterprise Tomography approach can control the Integration Lifecycle Management of Enterprise Clouds and increase the overall quality in an Enterprise Cloud according to free definable metrics while fulfilling requirements in a prioritized manner.

4 Related Works

The approach Enterprise Tomography driven Governance of Federated ERP in a Cloud is complementary to the research areas Application Lifecycle Management of VLBAs and governance of Semantic SOA respectively. In Semantic SOA there are dedicated procedures in alignment of semantic entities and semantic services [8]. The Enterprise Tomography approach generically unifies a set of ontology matching approaches and is primarily based on algorithms for genetic engineering known in Bio-Informatics [9,10,11,12]. The mathematical model of a family of matching algorithms for large data sets in genetic engineering is transformed to semantic matching and delta determination. The delta indicators can be interpreted as generic software metrics in a specific domain called semantic view. The software metrics are the decision basis in the governance procedure. Regarding metrics, service provisioning and consumption (dependency graph), business data as well as metadata is taken into consideration.

5 Conclusion

In this paper we have outlined the Federated EPR approach in the context of Enterprise Cloud Computing. It was discussed how FERP can increase scalability in a cloud. In addition we adumbrated the Integration Lifecycle Management of a Federated ERP network in a Cloud. With help of closed-loops the evolution of an shared Federated ERP system can be controlled according to cloud metrics, which are indicators calculated by the Enterprise Tomograph. The Enterprise Tomograph acts as a generic Delta-calculating search engine, which permanently crawls and observes the materialized cloud content. The search engine of the Enterprise Tomograph can be executed in delta mode as well as in full mode. With help of extractors for the Enterprise Tomograph we can have polymorphic search operator or delta operator which delivers the indicators as decision basis in the governance procedure.

References

1. Brehm, N., Marx Gòmez, J., Rautenstrauch, C.: An ERP solution based on web services and peer-topeer networks for small and medium enterprises. International Journal of Information Systems and Change Management (IJISCM) 1(1), 99–111 (2006)
2. Brehm, N., Lübke, D., Marx Gòmez, J.: Federated Enterprise Resource Planning (FERP) Systems. In: Saha, P. (Hrsg.) Handbook of Enterprise Systems Architecture in Practice, pp. 290–305. IGI Global, Hershey (2007)
3. Aalmink, J., Marx Gòmez, J.: Enterprise Tomography - an efficient approach for semi-automatic localization of integration concepts in VLBAs. In: Cruz-Cunha, M.M. (ed.) Social, Managerial and Organizational Dimensions of Enterprise Information Systems (2009) ISBN: 978-1-60566-856-7
4. Heuser, L., Alsdorf, C., Woods, D.: The Web-Based Service Industry - Infrastructure for Enterprise SOA 2.0, Potential Killer Applications - Semantic Service Discovery. In: International Research Forum 2008, Potsdam, SAP Research. Evolved Technologist Press (2008)
5. Heuser, L., Alsdorf, C., Woods, D.: Enterprise 2.0 - The Service Grid - User-Driven Innovation - Business Model Transformation. In: International Research Forum 2007, Potsdam, SAP Research. Evolved Technologist Press (2007)
6. Plattner, H.: A Common Database Approach for OLTP and OLAP using an In-Memory Column Database. In: International Conference on Management of Data. Proceedings of the 35th SIGMOD international conference on Management of data, Providence, Rhode Island, USA, pp. 1–2 (2009) ISBN: 978-1-60558-551-2
7. Epstein, A.: Parallel hardware architectures for the life science. Doctoral thesis. Delft University Press (2004)
8. Panchenko, O.: Concept Location and Program Comprehension in Service-Oriented Software. In: Proceedings of the IEEE 23rd International Conference on Software Maintenance: Doctoral Symposium, ICSM, Paris, France, pp. 513–514 (2007)
9. Tiun, S., Abdullah, R., Kong, T.E.: Automatic Topic Identification Using Ontology Hierarchy. In: Gelbukh, A. (ed.) CICLing 2001. LNCS, vol. 2004, pp. 444–453. Springer, Heidelberg (2001)
10. Haak, L., Brehm, N.: Ontologies supporting VLBAs; Semantic integration in the context of FERP. In: 3rd International Conference on Information and Communication Technologies: From Theory To Applications, ICTTA 2008, pp. 1–5 (2008)

11. Aalmink, J., Marx Gòmez, J.: Enterprise Tomography - an efficient Application Lifecycle Management approach supporting semiautomatic localization, delta-tracking and visualization of Integration Ontologies in VLBAs. In: Kumar, S., Bendoly, E., Esteves, J. (eds.) Frontiers of Research in Enterprise Systems, scheduled publication (2010)
12. Abels, S., Haak, L., Hahn, A.: Identification of common methods used for ontology integration tasks. Interoperability Of Heterogeneous Information Systems. In: Proceedings of the first international workshop on Interoperability of heterogeneous information systems, Bremen, Germany, pp. 75–78. ACM, New York (2005)

Composite Index of e-Business Strategy Readiness of the Enterprises in the Republic of Macedonia

Marjan Angeleski[1], Pece Mitrevski[2], and Margarita Janeska[1]

[1] Faculty of Economics, Gjorce Petrov bb, 7500 Prilep, Macedonia
[2] Faculty of Technical Sciences, Ivo Lola Ribar bb, 7000 Bitola, Macedonia
{marjan.angeleski,pece.mitrevski,margarita.janeska}@uklo.edu.mk

Abstract. The aim of this survey is focused on measuring the level of e-readiness of the enterprises in the Republic of Macedonia with an emphasis on the concept of "e-business strategy readiness".

The survey resulted in 348 responses from the Macedonian enterprises structured according to their economic activity (8 groups) and divided into 8 regions in accordance with Nomenclature of Territorial Units for Statistics – NUTS proposed by State Statistical Office of the Republic of Macedonia.

Based on this, we examine the indicators of the e-business strategy readiness index. This index is comprised of three core sub-indices: the level of adoption of ICT, the level of ICT usage, and the level of ICT strategy readiness. Furthermore, the e-business strategy readiness index and its composite sub-indices for the enterprises in the Republic of Macedonia have been calculated.

Keywords: composite index, e-business readiness, e-business strategy readiness.

1 Introduction

The concept of e-readiness and especially the concept of e-business strategic readiness are one of the most important indicators to show how efficiently a country could fight the competitiveness on the global market. By measuring these indicators the business entities will provide a way to realize their strengths and weaknesses in the process of digital restructuring and to see to the world trends, but also to realize the volume of its capacity for participation in the new global digital economy as well.

Assessment of the capacity of a certain nation for participation in the new economy is the first step in taking action in the ICT sector, whether it is an activity initiated by the government or the business sector. E-readiness assessment should be an incentive for improving the capacities of countries with a high index of e-readiness and their improvement, especially among those with lower index e-readiness, and the entire in order to be competitive in the global economy. On the other hand, many business strategies begin with examinations or assessments of the current state of the enterprises for highlighting the relative strengths and weaknesses and opportunities.

During the development of business strategies that included information and communication technology, the assessment of the current situation represents the baseline for the extent of the changes that are indispensable to take in the future. Starting from these assumptions, in this research was made analysis of factors affecting composite index of information and communication readiness for investment, use and implementation of e-business strategies in the operations of business entities in Macedonia.

The term e-readiness in a broad context can be defined as the capacity of a nation to participate in the digital economy or the ability of a nation to establish ICT connection with the rest of the world. E-readiness measures the level of ICT adoption and the capability for ICT use in all the fields related to socio-economic functioning.

The purpose of this survey is not allocated on the common definition for e-readiness but the focus is initially placed more precisely upon the e-business strategic readiness of the business entities in the Republic of Macedonia. The concept of e-business strategy readiness enables for the assessment of the level of implementation and ICT usage among business entities, as well as the readiness for incorporating e-business concepts in their business strategy and policy.

From these reasons, it is very important for every country to know the volume of its capacity for participation in the digital economy, not only for the purposes of the statistics, but from the aspect of comparison with other countries and undertaking appropriate measures for ICT expansion which will improve the business climate and attract foreign investors.

2 Survey Instrument and Methodology

2.1 Indicators of the Composite Index of e-Business Strategic Readiness

The index of e-business strategic readiness has been composed of three groups of indicators: ICT adoption, ICT use and ICT strategic readiness. The first two are defined by the Joint Research Centre of the European Commission. But, one of the main goals in this survey was to include one more group of indicators that refer to the ICT strategy readiness of the companies. Thus, taking into consideration the importance from strategic way of investment and special use of information and communication technology as the main pillar and basis for development of e-business concept, here, for the first time, a third group of indicators is introduced and applied on the level of business entities. This group of indicators refers to strategic aspects of electronic business. The indicators were chosen in order to correspond with the main components that the management of the business entity should know while defining and formulating the e-business strategy. In choosing this group of indicators from the large number of alternative indicators, a special care has been taken of those that show representation of these problems and crucially sublime the problems as a whole, theoretically described in the strategic e-business management. In addition, a care has been taken of the questions. They are simple, structural, easily understandable, closed and they should give answers to the crucial questions, essential while giving general evaluation for the state of this field in the Republic of Macedonia.

Therefore, the composite index of e-business strategy readiness has been composed by following indicators:

I. Adoption of ICT: basic indicators

(a1) Percentage of enterprises that use Internet;

(a2) Percentage of enterprises that have web/home page;

(a3) Percentage of enterprises that use at least two 2 security facilities at the time of the survey;

(a4) Percentage of total number of persons employees using computer with their normal work;

(a5) Percentage of enterprises having broadband connection to internet;

(a6) Percentage of enterprises with LAN and using an Intranet or Extranet.

II. Use of ICT: basic indicators

(b1) Percentage of enterprises that have purchased products/services via the internet, EDI or any other computer mediated network where these are >1% of total purchases;

(b2) Percentage of enterprises that have received orders via the internet, EDI or anyother computer mediated network where these are >1% of total turn over;

(b3) Percentage of enterprises whose IT systems for managing orders or purchases are linked automatically with other internal IT systems;

(b4) Percentage enterprises whose IT systems are linked automatically to IT systems of suppliers or customers outside their enterprise group;

(b5) Percentage of enterprises with Internet access using the internet for banking and financial services;

(b6) Percentage of enterprises that have sold products to other enterprises via a presence on specialized internet marketplaces.

III. ICT strategic readiness: basic indicators

(c1) percentage of business entities that have had clear image for potential e-business gains, applications, trends, possibilities and models that could be used in their firm;

(c2) percentage of business entities that have surveyed strategic initiatives of some of their competitors and other participants in the sector and have also considered how the implementation of these initiatives could affect their competitiveness;

(c3) percentage of business entities that have been familiar with the characteristics of some types of e-business strategies;

(c4) percentage of business entities that have had employed ICT experts in their firm;

(c5) percentage of business entities that in the vision and mission of their have been emphasized separate segments of e-business;

(c6) percentage of business entities that have been familiar with the legal framework which regulates the subject matter connected with e-business in our country.

2.2 Sample

The aimed group of survey was the active business entities in the Republic of Macedonia. Namely, these business entities were classified into two groups, and the survey was performed in two phases, since the calculation of the composite indexes of e-business readiness was performed according to economic activities and the result is common, that is, it refers to all business entities in the Republic of Macedonia.

The business entities which meet the criterion of "large companies", according to their total incomes realized in 2006 and published in the edition "The largest 200", were classified in the first group. "The largest" companies from these 200 made a total annual income of 494 million euros, while the "least" made a total annual income of 6 million euros. In the second group, there are the business entities that made total annual income lees than 6 million euros, classified according the National classification of economic activity in 7, or rather 8 groups (all business entities, whose economic activity does not belong to no one of 7 defined activities, are included in the eighth group under the name of other activities) and divided into 8 regions in accordance with Nomenclature of Territorial Units for Statistics – NUTS proposed by State Statistical Office of the Republic of Macedonia.

The survey was performed from June 2007 to June 2008 principally by e-mail or verbal communication with liable persons of companies covered by this survey. The survey resulted in 348 responses from the Macedonian enterprises from different activities.

The schedule of the polled business entities and their sectors and regions is presented in Table 1.

Table 1. Structural schedule of all polled business entities

Sectors and reagions	Pelagonia	Vardar	Northeastern	Southwestern	Skopje	Southeastern	Polog	Eastern	Total
I. Manufacturing	6	5	3	8	21	9	5	6	63
II. Construction	3	1	1	1	9	1	1	1	18
III. Wholesale and retail trade	18	13	10	17	59	13	16	14	160
IV. Finance	3	1	1	1	8	1	2	1	18
V. Transport, storage and comunication	2	2	2	3	15	2	3	2	31
VI. Catering and tourism industry	3	1	1	1	8	1	1	1	17
VII. Motion picture, video, radion and TV activities	2	2	1	2	8	2	2	2	21
VIII. Other industries	1	1	1	2	9	2	2	2	20
Total	38	26	20	35	137	31	32	29	348

3 Results and Discussion

The results received from the poll are divided in groups of indicators in Table 2, Table 3 and Table 4.

Table 2. ICT adoption

Indicators / Sector	a1		a2		a3		a4		a5		a6	
	affirmative answers		affirmative answers		affirmative answers		affirmative answers		affirmative answers		affirmative answers	
I. Manufacturing	43	68,25%	24	38,10%	31	49,21%	28	44,44%	48	76,19%	21	33,33%
II. Construction	11	61,11%	6	33,33%	8	44,44%	5	27,78%	12	66,67%	7	38,89%
III. Wholesale and retail trade	124	77,50%	71	44,38%	89	55,63%	96	60,00%	114	71,25%	37	23,13%
IV. Finance and busines activities	16	88,89%	16	88,89%	16	88,89%	16	88,89%	14	77,78%	9	50,00%
V. Transport, storage and comunication	26	83,87%	14	45,16%	21	67,74%	19	61,29%	19	61,29%	14	45,16%
VI. Catering and tourism Industry	9	52,94%	9	52,94%	8	47,06%	9	52,94%	9	52,94%	4	23,53%
VII. Motion picture, video, radio and TV activities	20	95,24%	15	71,43%	17	80,95%	17	80,95%	17	80,95%	5	23,81%
VIII. Other industries	10	50,00%	4	20,00%	6	30,00%	8	40,00%	11	55,00%	4	20,00%
Total	**259**	**74,43%**	**159**	**45,69%**	**196**	**56,32%**	**198**	**56,90%**	**244**	**61,25%**	**101**	**29,02%**

Table 3. ICT use

Indicators / Sector	b1		b2		b3		b4		b5		b6	
	affirmative answers		affirmative answers		affirmative answers		affirmative answers		affirmative answers		affirmative answers	
I. Manufacturing	4	6,35%	2	3,17%	27	42,86%	3	4,76%	36	57,14%	4	6,35%
II. Construction	1	5,56%	1	5,56%	2	11,11%	1	5,56%	11	61,11%	2	11,11%
III. Wholesale and retail trade	12	7,50%	5	3,13%	84	52,50%	5	3,13%	89	55,63%	20	12,50%
IV. Finance and business activities	2	11,11%	2	11,11%	10	55,56%	2	11,11%	15	83,33%	3	16,67%
V. Transport, storage and comunication	2	6,45%	1	3,23%	5	16,13%	1	3,23%	11	35,48%	5	16,13%
VI. Catering and tourism Industry	2	11,76%	3	17,65%	6	35,29%	1	5,88%	11	64,71%	6	35,29%
VII. Motion picture, video, radio and TV activities	1	4,76%	2	9,52%	0	0,00%	0	0,00%	9	42,86%	5	23,81%
VIII. Other industries	1	5,00%	1	5,00%	1	5,00%	0	0,00%	4	20,00%	1	5,00%
Total	**25**	**7,18%**	**17**	**4,89%**	**135**	**38,79%**	**13**	**3,74%**	**186**	**53,45%**	**46**	**13,22%**

Table 4. ICT strategic readiness

Indicators / Sector	c1		c2		c3		c4		c5		c6	
	affirmative answers		affirmative answers		affirmative answers		affirmative answers		affirmative answers		affirmative answers	
I. Manufacturing	32	50,79%	9	14,29%	34	53,97%	16	25,40%	9	14,29%	9	14,29%
II. Construction	7	38,89%	2	11,11%	5	27,78%	2	11,11%	1	5,56%	2	11,11%
III. Wholesale and retail trade	98	61,25%	24	15,00%	99	61,88%	11	6,88%	14	8,75%	18	11,25%
IV. Finance and business activities	16	88,89%	11	61,11%	16	88,89%	14	77,78%	11	61,11%	12	66,67%
V. Transport, storage and comunication	22	70,97%	14	45,16%	16	51,61%	11	35,48%	9	29,03%	15	48,39%
VI. Catering and tourism Industry	11	64,71%	9	52,94%	8	47,06%	2	11,76%	2	11,76%	2	11,76%
VII. Motion picture, video, radio and TV activities	10	47,62%	10	47,62%	6	28,57%	9	42,86%	4	19,05%	4	19,05%
VIII. Other industries	7	35,00%	3	15,00%	2	10,00%	1	5,00%	0	0,00%	1	5,00%
Total	**203**	**58,33%**	**82**	**23,56%**	**186**	**53,45%**	**66**	**18,97%**	**50**	**14,37%**	**63**	**18,10%**

From the Chi square testing of hypothesis for the dependence between measured indicators and the sector of the business entities in this segment, a conclusion could be made that a part of the factors is in statistic dependence on the sector of the business entities, while the other part is not.

Therefore, the testing hypothesis could lead to a resume that there is a difference between possessing web page and the sector of the company ($\chi^2 = 27{,}510 > 14{,}067$, $df = 7$, $p < 0{,}0003$, $\alpha = 0{,}05$). Also, whether one business entity uses a broadband Internet connection or some other type of it, does not depend on which sector it belongs to, but this is produced by other factors ($\chi^2 = 8{,}717 > 14{,}067$, $df = 7$, $p < 0.274$, $\alpha = 0{,}05$).

From Chi square testing the indicators connected with ICT use, two hypotheses could be separated. Therefore, the hypothesis for dependence between using Internet approach for separate bank and financial services in relation to the sector of the business entities ($\chi^2 = 22{,}363 > 14{,}067$, $df = 7$, $p < 0{,}002$, $\alpha = 0{,}05$) and the hypothesis which shows that there is a statistic dependence (but on a higher significant level of 0,1) between the sector of the business entity and the fact whether it sells products to other firms via specialized electronic markets ($\chi^2 = 13{,}602 > 12{,}017$, $df = 7$, $p < 0{,}059$, $\alpha = 0{,}1$).

The results of the Chi square testing for ICT strategic readiness of business entities from different sectors, show that one part of the entities do not have equally clear image for the potential e-business gains, applications, trends, possibilities and models that could be applied. In other words, there is dependence between the variable that describes these categories and the sector of the business entities themselves ($\chi^2 = 19{,}540 > 14{,}067$, $df = 7$, $p < 0{,}007$, $\alpha = 0{,}05$)). This indicates that the business entities from particular sectors are more informed compared to business entities from other sectors. Also, the testing has shown that there is a high statistic significance at a level of 0,05, i.e. the "awareness" depends on the sector of the business entity ($\chi^2 = 39{,}146 > 14{,}067$, $df = 7$, $p < 0{,}0001$, $\alpha = 0{,}05$).

4 Calculation of the Composite Index of e-Business Strategic Readiness

Composite index of e-business strategic readiness, as has been previously noted, is composed of three sub-indices (composite index of e-business strategic readiness and its components is shown in Figure 1). The main goal is to aggregate all those sub-indices under one common index. Initially, the sub-indices of ICT adoption, ICT use and ICT strategic readiness are calculated, and then the composite index of e-business strategic readiness is assessed. Linear aggregation has been used and all the calculations were carried out using the XLSTAT software tool.

Most of the composite indicators are based on equal weighting, i.e. all the variables involved are equally important for the composite index. But, weighting can be taken into account so to reflect the quality of statistical data. In this case, the determination of the weights can be done by multivariate PCA/FA analysis (this method is described by the Joint Research Centre of the European Commission).

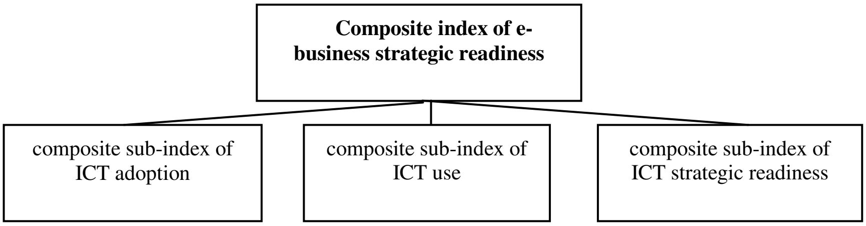

Composite index of e-business strategic readiness
composite sub-index of ICT adoption
composite sub-index of ICT use
composite sub-index of ICT strategic readiness

Fig. 1. Conceptual framework for evaluation of the composite index of e-business strategic readiness and its integral sub-indexes

4.1 Calculation of the Composite Sub-indices

The composite sub-index CI_s^{ict-a} of ICT adoption for specific sector s is a linear sum of the products of the weights w_{ak} ($0 \le w_{ak} \le 1$ and $\sum_{i-1}^{k} w_{ak} = 1$) and the normalized values of the sub-indicators of this sector a_{sk}. This means that

$$CI_s^{ict-a} = \sum_{i=1}^{k} w_{ak} a_{sk} \tag{1}$$

Following the steps of calculating the weights of each of the indicators first is calculated the correlation matrix. The values of the correlation matrix are given by the Table 5.

The marked values in the correlation matrix are significant at the level of $\alpha = 0{,}05$ (two-tailed test). This means that there should be a relatively large number of significant values for using this method[1].

Table 5. Correlation matrix of the ICT adoption variables

	a1	a2	a3	a4	a5	a6
a1	1	**0,738**	**0,928**	0,703	**0,773**	0,423
a2	**0,738**	1	**0,922**	0,663	0,581	0,411
a3	**0,928**	**0,922**	1	0,673	0,683	0,533
a4	0,703	0,663	0,673	1	0,461	-0,152
a5	**0,773**	0,581	0,683	0,461	1	0,283
a6	0,423	0,411	0,533	-0,152	0,283	1

In order to provide stronger arguments for the existence of correlation Bartlett's sphericity test has been performed ($\chi^2 = 50{,}575 > 24{,}996$, $p < 0{,}0001$, $\alpha = 0{,}05$).

The next step is to calculate the eigenvalues and to determine the number of factors that should be retained in further calculations. The eigenvalues are given in Table 6.

Table 6. Eigenvalues for indicators of ICT adoption

	F1	F2	F3	F4	F5	F6
Eigenvalue	4,042	1,149	0,518	0,225	0,065	0,000
% variance	67,368	19,151	8,635	3,755	1,089	0,003
Cumulative %	67,368	86,519	95,154	98,908	99,997	100,000

According to the criterion of explained variation from the previous table, we can conclude that it is necessary to keep the first two factors as they cumulatively explain more than 80%, or more precisely 86.519% of the total variation.

[1] If there is no correlation between the indicators, the PCA/FA method can not be used for determining the weight coefficients and then it is necessary to use other methods.

The next step is the rotation of factors' loadings, retaining only the first two factors. The percentage of variation explained after rotation remains the same. Factors' loadings after Varimax rotation are shown in the Table 7.

Table 7. Factor loadings after Varimax rotation

	F1	F2
a1	0,919	0,243
a2	0,865	0,246
a3	0,920	0,354
a4	0,874	-0,400
a5	0,759	0,208
a6	0,205	0,960

The last step is the construction of the weight coefficients of the Factors' loadings matrix after Varimax rotation. In this case, we have used the approach proposed by Nicoletti G., Scarpetta S., Boylaud O. Namely. First, each factor loading is squared. Then, those with the greatest values are chosen. Values are divided by the sum of the largest factor loadings, so the weight coefficients of each indicator are obtained. Consequently, the weights w_{ak} determined using PCA/FA methodology are shown in the Table 8.

Table 8. The weights for indicators of ICT adoption obtained using PCA/FA

w_{a1}	0,180
w_{a2}	0,159
w_{a3}	0,180
w_{a4}	0,162
w_{a5}	0,123
w_{a6}	0,196

By analogy, the weights of composite sub-indices of ICT use and ICT strategic readiness are calculated. The weights are shown in Table 9 and Table 10, respectively.

Table 9. The weights for indicators of ICT use obtained used PCA/FA

w_{b1}	0,115
w_{b2}	0,185
w_{b3}	0,172
w_{b4}	0,175
w_{b5}	0,163
w_{b6}	0,190

Table 10. The weights for indicators of ICT strategic readiness obtaind used PCA/FA

w_{c1}	0,177
w_{c2}	0,127
w_{c3}	0,142
w_{c4}	0,171
w_{c5}	0,199
w_{c6}	0,185

Using the previously defined models, the composite sub-indices and the composite index of e-business strategy readiness by sectors are given in Table 11.

Table 11. The values of the composite sub-indexes and overall composite index of e-business strategy readiness

Sectors	CI_s^{ict-a}	CI_s^{ict-u}	CI_s^{ict-sr}	CI_s^{ebsr}
I. Manufacturing	0,447	0,200	0,283	**0,310**
II. Construction	0,425	0,166	0,173	**0,255**
III. Wholesale and retail trade	0,542	0,224	0,258	**0,341**
IV. Finance and business activities	0,726	0,316	0,738	**0,593**
V. Transport, storage and comunication	0,557	0,135	0,464	**0,385**
VI. Catering and tourism Industry	0,430	0,290	0,314	**0,345**
VII. Motion picture, video, radio and TV activities	0,682	0,138	0,331	**0,384**
VIII. Other industries	0,301	0,066	0,113	**0,160**

Figure 2 shows the total composite index of e-business strategic readiness by sectors.

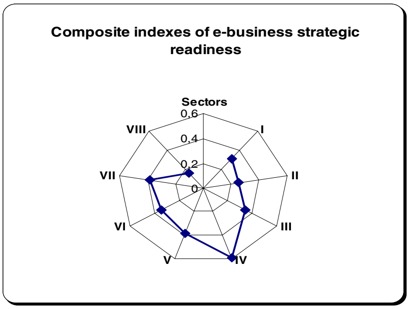


Fig. 2. Composite indexes of e-business strategic readiness

The performed calculations generally show that the quantitative scores for ICT adoption are relatively higher than the indices of ICT use and ICT strategic readiness. The performed tests for the Spearman's rank correlation coefficient show that there is no quantitative reconciliation between the first and the second group of indicators, that is, the Spearman's rank correlation coefficient for these groups of indicators is 0,357, which means that at a level of importance $\alpha = 0{,}05$ the null hypothesis for the existence of significant correlation should not be neglected. This is not the case here, although the theoretical assumptions would be that there is a significant correlation between the ranges of the first and the second group of indicators. The non-existence of significant correlation primarily lies in the inappropriate way of using the ICT that the business entities have invested in. The coefficient of range correlation between the indicators of ICT investment and ICT strategic aspects is interesting. Namely, there is a significant dependence between the ranges of these two groups of indicators (Spearman's rank correlation coefficient of 0,881), which means that the null hypothesis for the non-existence of significant correlation should be neglected (level of importance $\alpha = 0{,}05$). These results lead to a conclusion that the largest part of ICT investment has been strategically premeditated. But, the Spearman's rank correlation coefficient of 0,381 between the ranges of business entities for ICT use and ICT strategic aspects shows that there is no statistically important range correlation, i.e. ICT is not used with previously premeditated strategic concepts.

The performed analyses could also state that business entities from certain sectors are positioned beyond the level of their potential capabilities – although the business entities have invested in ICT, they neither used the capacities of ICT investment nor strategically used such investments.

The performed comparison of calculated indices for e-business readiness with those of the EU countries could lead to a conclusion that although this index is lower in the Republic of Macedonia, it still hasn't fallen behind to a great extent – on the contrary, it is followed by the index of some countries in eastern Europe (Table 12).

Table 12. The values of the composite indexes for e-business readiness of the EU countries (JRC 2006) and Macedonia (author's survey, June 2007-June 2008)

Countries	AT	BE	BG	CY	CZ	DK	EE	FI	FR	DE	GR	HU	IS	IE	IT	LV	LT	LU	NL	NO	PL	PT	RO	SK	SI	SP	UK	EU27	MK
ICT adoption	0,680	0,717	0,430	0,512	0,601	0,757	0,699	0,781	0,691	0,701	0,554	0,488	0,760	0,644	0,605	0,454	0,514	0,679	0,726	0,713	0,521	0,523	0,324	0,574	0,634	0,632	0,684	0,639	0,514
ICT use	0,307	0,284	0,077	0,192	0,228	0,414	0,234	0,309	0,301	0,330	0,267	0,122	0,324	0,332	0,240	0,137	0,212	0,270	0,352	0,343	0,174	0,193	0,110	0,220	0,217	0,229	0,279	0,265	0,192

5 Conclusion

The evaluation of the index of e-business strategic readiness of the enterprises is very important towards encouraging their innovation and competitiveness. The performed calculations generally show that the quantitative scores for ICT adoption are relatively higher than the indices of ICT use and ICT strategic readiness. In addition, it can be concluded that the highest index of e-business strategic readiness of 0.593 have the enterprises performing financial and business activities, while the highest index of e-business strategic readiness of 0.255 have the companies with economic activity in the construction sector.

The questions and dilemmas left behind give wide opportunities for future research, given that there is a need of continuous annual survey that will lead to calculation of the index of e-business strategic readiness in a longer period of time. This will be an excellent base for predictions, but also comparisons of the indices in the Republic of Macedonia with other European countries in the period ahead.

References

1. Handbook on Constructing Composite Indicators: Methodology and User Guide. OECD & Joint Research Centre (2008)
2. Spremić, M., Strugar, I.: Strategic IS Planning Practice in Croatia: Organizational and Managerial Challenges. International Journal of Accounting Information Systems 3(3), 183–201 (2002)
3. Nardo, M., Saisana, M., Saltelli, A., Tarantola, S.: Tools for Composite Indicators Building. Joint Research Centre (2005)
4. Pennoni, F., Tarantola, S., Latvala, A.: The European e-Business Readiness Index. Joint Research Centre (2003–2007)

Performance Evaluation of a New Approach for Automatic Question Production

Mile Jovanov and Marjan Gusev

Institute of Informatics, FNSM, Gazi Baba b.b., 1000 Skopje
{mile,marjan}@ii.edu.mk

Abstract. Nowadays, e – testing is an often used method for evaluation in the process of learning. In this paper, we discuss the e – testing problem of creating large question set that will reflect the knowledge of some domain. A new model of E – testing is introduced with a proposal of a new solution to the problem of creation a large question set for a given domain. Then, we present a methodology for comparison of the results and the contribution of the new model and realization on the automated creation of large number of questions, and we evaluate the quality and the vulnerability of the question set, as well. It is shown that the new model increases the speed of question production by more 10 times.

Keywords: Semantic Web, semantic web technologies, ontology, OWL, e - testing, question set.

1 Introduction

Semantic Web is an evolving extension of WWW, in which the meaning of information and services on web are defined in such way to allow computers to understand and satisfy human requests using the web content. The goal is to develop standards and technologies designed to help machines understand more information on the web so that they can support richer discovery, data integration, navigation, and automation of tasks.

Semantic Web is an attempt to address the initial goal of the web enabling automation. Short term goal of the Semantic Web is interoperability, and long term goal is to make computers work on our behalf instead of using them like tools [1].

OWL, as one of the developed technologies, is the language for description of ontologies. OWL document describes an existing ontology.

Semantic Web technologies can be employed in many areas of computer science. In this paper we use OWL documents in area of e-learning, particularly e-testing.

E-learning is a process of education in electronic form through Internet network or the Intranet with the use of management system for education. Evaluation is important step in learning and e-learning process.

The process of electronic evaluation of students is referred to as e-testing, web testing, online quiz, etc.

An e-test consists of set of questions that could be: multiple choice, true/false, ordering, matching, drag and drop, essay, etc.

The test could have a time limit or not, even more, every question could be time limited with different time. The question set could be predetermined or the questions could be given depending on the previous answers of the student.

Advantages of e-testing over regular testing are numerous. For example, among the possibilities offered by “Moodle” platform are the following [2]:

- Teachers can define a database of questions for re-use in different quizzes;
- Questions can be stored in categories for easy access, and these categories can be "published" to make them accessible from any course on the site;
- Quizzes are automatically graded, and can be re-graded if questions are modified;
- Quizzes can have a limited time window outside of which they are not available;
- At the teacher's option, quizzes can be attempted multiple times, and can show feedback and/or correct answers;
- Quiz questions and quiz answers can be shuffled (randomized) to reduce cheating;
- Questions allow HTML and images;
- Questions can be imported from external text files;
- Quizzes can be attempted multiple times, if desired;
- Attempts can be cumulative, if desired, and finished over several sessions.

E-testing allows evaluation of large number of students which can be very helpful in institutions where student-teacher ratio is high.

Additional features offered by e-testing provide learning manager (i.e. teacher) a tool for student self evaluation in the process of learning. Also, large set of different type of questions permit more accurate evaluation of the student. The possibility of re-grading the quizzes after modification of some question(s) offers flexibility and quick recovery if some mistake or inaccuracy in given questions is noticed.

2 Weaknesses of e-Testing

E-testing as well as regular testing has more weaknesses. One of major ones is collecting (printing, saving, etc…) questions by the students and sharing the copies among them. This can happen when the test is set to be taken by the student in unattended (and unsecured) environment, and also when the testing is performed in classroom where students are proctored by someone.

If the environment (web or application) of the test is not secure enough, possibility of cheating through going forward and backward, delaying the time, accessing other recourses is also present.

But, most important issue is question database. The questions included in e-tests can be taken from question database. With every test a part of the database is exposed. If students can save this questions they can quickly have the question database (or main part of it) so after that results from the testing will not illustrate the knowledge of the student on the subject, but just on the database.

When dealing with students that have more computer skills (IT students) one should be aware that they could try to attack the database directly using SQL injection, URL manipulation, buffer overflow, remote command execution, weak authentication and authorization, etc. [4]

When students have the questions in electronic format then if access to other applications and processes on the computer where the e-testing occurs isn't protected, students may simply search thru the list of questions (as simple as option "Find"), and just see the right answer of the given question.

Feeling comfortable about test security usually comes down to feeling comfortable that (a) the person whose name is associated with the test is indeed the person who took the test and (b) the students were not exposed to the test items before taking the test. If that comfort isn't provided through an honor code, it has to be established through the testing procedures. [5]

But, the main question is if there is a way to discourage students to make a collection of the questions from the set of questions that the teachers have. One solution already implemented in some e-testing environments is randomizing the question order and the order of answers (for example, [3]). It makes the printouts a lot less useful.

Creating larger question banks and giving tests with random subsets is also an effective strategy. If students can only print a small number of questions at a time, they will need to view the test again and again, and then sort the questions to eliminate duplicates. In this way, memorizing the questions will be rather difficult.

Very clear observation made by many researchers (for example, [7]) is that creating a question database is time-consuming. This is the task that nowadays should be done by teachers. Creating only a minimal set of questions could take more than 10 hours work per week. [6]

The question that remains open is how to create a large set of questions. This is the question of interest in this paper.

3 How to Create a Large Set of Questions for e-Testing?

One direction in which one could look for the solution is the existence of large community of teachers that can use same standard for produced questions. For example, Advanced Distributed Learning (ADL) has offered Sharable Content Object Reference Model (SCORM) which integrates a set of related technical standards, specifications, and guidelines designed to meet SCORM's high-level requirements — accessible, reusable, interoperable, and durable content and systems. SCORM content can be delivered to learners via any SCORM-compliant Learning Management System (LMS) using the same version of SCORM. [6]

In this way, large sets could be easily created but only in languages that are massively spoken and only on more common topics. Additionally, great effort should be put in division of questions in categories and subcategories.

The other direction that we propose is use of software for automatic creation (generation) of the questions. The proposed software should be able to produce a large set of questions using files that contain knowledge of a certain domain. These files should contain knowledge in "non-linear" way, difficult to be memorized by the students. The application should offer different structures of questions and possibility to change the fixed text of the question.

4 A Model for Automatic Question Production for e-Testing Systems

The model that we propose is given on Figure 1.

Semantic web technology, OWL (web ontology language) in particular, offers a way of "non-linear" description of knowledge. Nowadays, OWL files describing ontologies are produced every day for many specific domains. These files are used as sources for the produced software built on the model. The software extracts the knowledge from the file by parsing and then produces a large number of questions concerning the described domain. The questions can be of different type, but more preferably multi-choice and true-false questions, easy for computer grading.

Produced questions can be used in two ways. First option is, an other part of the software to generate the test by choosing a random subset of the questions. The test can be used to grade a student (or more students). Second option is to export this

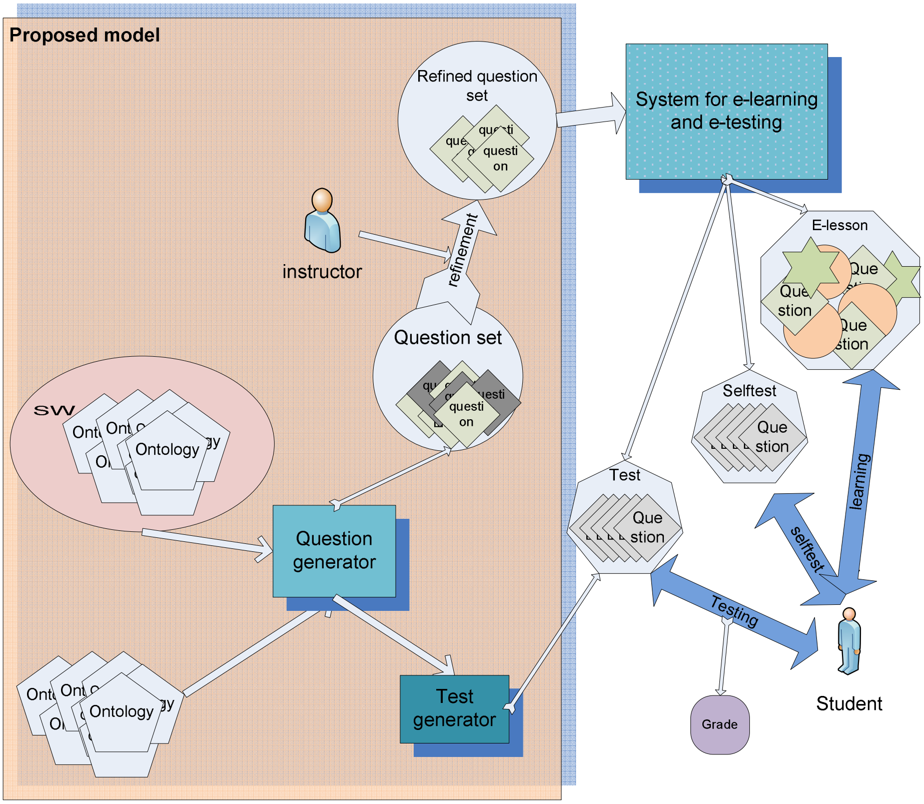


Fig. 1. A model for automatic question production for e-testing systems

questions in some format (preferably XML) and to store them. This option gives additional possibility for the set to be checked by qualified instructor in order to make corrections to some of the questions (syntax and/or semantic) or to completely reject some. Such refined question set can be used in any Learning Content Management System that allows e-testing, self-testing and/or e-lessons.

The process of question generation consists of phase in which the knowledge is extracted from the input ontology, and the phase of question generation.

In the first phase, the document is parsed, and the data structures containing detected concepts (classes, properties) are created.

In the second phase, using the elements of the mentioned structures, different form of questions are created. In the software that we produced based on the proposed model there are 27 different types of multi-choice and true-false questions (such as, questions about relations between classes, properties, characteristics of classes and properties). Different type of sub algorithm decides on the false answers that will be offered in the multi-choice questions, to mach the question itself. With exhaustive search every possible question is created.

5 Methodology to Evaluate the Model

Proposed model tries to solve the problem of the question set vulnerability. Therefore, the following characteristics are evaluated:

- Question production speed,
- Good question formulation,
- Solvability of the questions.

Question production speed is key criteria for measuring the quality of the proposed solution, as the main goal of the solution is fast production of questions. It is measured through the time interval for creating a question (or fixed number of questions), the time interval for checking a question (or fixed number of questions), which sums up to the time for producing a question. The result is compared to the time for manual production of a question.

Good question formulation as a quality is measured by counting the rejected and fixed questions in the process of question checking (refinement phase) in both ways of production.

Solvability of the question represents "the possibility" for the question to be solved by the student. In reality, there are questions that can be solved by almost anyone, and as opposite, questions solvable by very small number of students. Coefficient of question solvability is calculated for every question using:

$$k_1 = \frac{\sum_{i=1}^{t} \frac{1}{t} p_i - \sum_{i=1}^{f} \frac{1}{2f} q_i}{N} \tag{1}$$

where t represents the number of true options, f – number of false options, p_i – number of students that have chosen the i-th true option, q_i – number of students that have chosen the i-th false option, and N – total number of students that had the possibility to answer the question.

A coefficient of answering the question is also calculated by:

$$k_2 = \frac{n}{N} \tag{2}$$

where n – is the number of students that have tried to answer the question, and N – total number of students that had the possibility to answer the question. It should be stated that every inaccurately answered question gives negative points to the final score of the student, so some of them decide not to answer some question.

6 Comparative Analysis of the Results

The software that we use in testing the model performance is "OWL_Question_generator". It is produced, as visual application, based on the model in Microsoft Visual Studio C++ 2005 Express Edition. It parses the OWL document on input and stores the extracted knowledge in various data structures. Then, using different algorithms generates different forms of multi choice questions. Questions are exported in suitable XML format.

We compare the performance of this model to the existing solution of manual production of questions. The results for the creation and checking (refinement) of the questions are gained through experiments done by 8 qualified instructors on the topic of Object and Visual Programming. The result about solvability of the questions are calculated from the results of the exam given to the students taking the course Object and Visual Programming.

Table 1 shows the results for the manual production of questions. Given that average time for production of question, the calculated question production speed is 0,1814 questions/min.

Table 1. Estimated time in the process of manual question production

	Average time in minutes	**Standard deviation**
Question creation	4,131	1,086
Question checking	1,381	1,068
Total time for question production:	**5,512**	

Table 2 shows the results for the automatic creation and manual checking of questions. Given that average time for production of question it is calculated that question production speed is 1,944 questions/min.

Table 2. Estimated time in the process of automatic question production

	Average time in minutes	**Standard deviation**
Question creation	0,0004	~0
Question checking	0,514	0,292
Total time for question production:	**0,5144**	

According to the previous result, we may conclude that even when the process of manual refinement of question set is included in the question production, the new model offers ***almost 11 times faster production***.

If we consider ***the good question formulation*** according to the results in the process of manual production of questions 39,88% of the questions were repaired (changed) and 7,14% were rejected. On the other hand, in the process of automatic production 2,44% were repaired and 0,35% rejected. So, in both cases (repairing or rejecting) process of automatic production shows ***over 16 times better results***.

Solvability of questions is calculated on every question in both sets by giving the questions to large number of students. The gained interval for the coefficient of solvability in both cases is [-0.3, 1]. Table 3 defines the boundaries of “classes” of solvability.

Table 3. Defined boundaries of the classes of solvability

Class	Interval of coefficient k1
1. “very hard to solve“ question	[-0,3; -0,04]
2. “hard to solve“ question	(0,04; 0,22]
3. “standard solvable” question	(0,22; 0,48]
4. “easy to solve“ question	(0,48; 0,74]
5. “very easy to solve“ question	(0,74; 1]

Figure 2 and Figure 3 show that produced questions in both ways are distributed in the 5 classes of solvability with no significant differences.

In the case of automatic production of questions slightly greater solvability, but more important in both cases there is non-uniform but good distribution among classes.

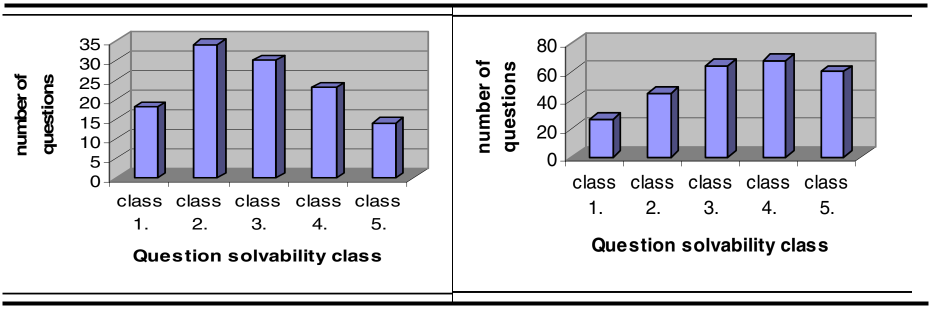


Fig. 2. Column charts showing the number of questions per class of solvability, produced manually (on left) and automatically (on right)

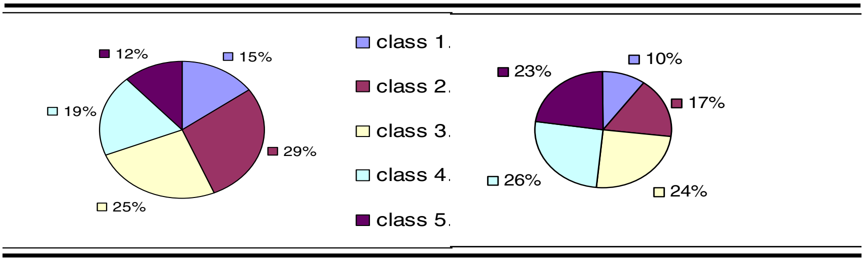


Fig. 3. Bar charts showing the percentage of questions per class of solvability, produced manually (on left) and automatically (on right)

Figure 4 represents coefficient of answering the question by classes. Here, the coefficient is in the interval [0; 1], and the five presented classes are [0; 0,2], (0,2; 0,4], (0,4; 0,6], (0,6; 0,8], (0,8; 1]. It can be concluded that students more bravely were answering the automatically produced questions. This is, probably, due to the fact that for automatically produced questions there is finite number of formulations of the questions.

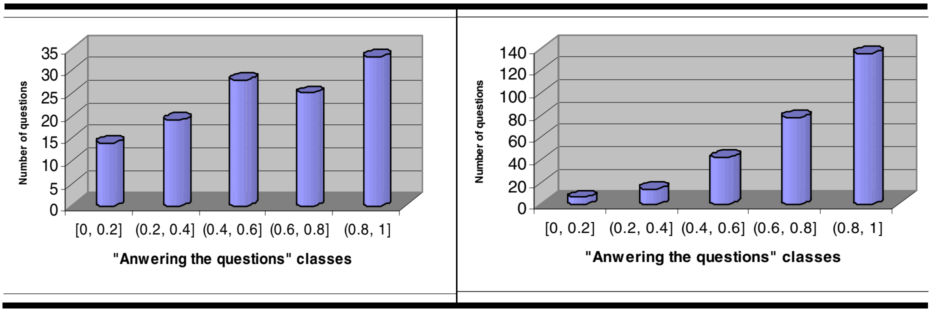


Fig. 4. Column charts showing the number of questions per class of answering the question, produced manually (on left) and automatically (on right)

Additional qualities offered by the new model are:

- Creating the questions unmistakably
- Form of question storage

Creating the questions unmistakably is an important quality which can be offered by any (well done) software in a process versus a process done, or partly done by human. In our case this quality depends on the produced software based on the model and on the ontology used as input in the software.

Form of question storage can have a great effect on the vulnerability of question set. Software based on presented model, can test the student even without a stored question set, because the questions can be produced in the same moment. In this case the advantage of this model is obvious, because in this way the knowledge is coded in the ontology, not in the question set. So, someone who tries to game the system can only get the ontology, but if she learns all the concepts and relations in it, she will have the necessary knowledge.

However, if we decide to use this approach we well have to sacrifice the possibility to store the questions in a database. In this case the testator will not be able to check the created questions and to select just part of them as a pool for testing.

On the other hand, even if we decide that we need to store the questions (to have possibility to check them) there is still an advantage because the main goal of a large question set is achieved.

7 Conclusion

The problem of vulnerability of the question set in e-testing systems motivated our research. In this paper we presented a performance evaluation of a new model for automatic question production that uses Semantic Web ontology (OWL document) as input. The model allows very fast production of large question set. Even with the additional checking of produced questions the production speed is 10 times bigger than in the process of manual production. Good results of the model are also shown on “good question formulation” quality. We showed that the set of automatically produced questions doesn’t significantly defer from the set of manually produced ones, in the sense of question solvability. So, the presented model could be used in the process of creation of questions for e – testing purposes.

References

1. Lasila, O.: Towards the semantic web. Presentation. W3C Semantic Tour, London (2003)
2. Moodle, Features, Quiz module,
   `http://docs.moodle.org/en/Features#Quiz_Module`
   (last accessed July 2009)
3. Gusev, M., Armenski, G.: E-Learning realized by E-Testing. In: Proceedings of the 2nd Conference on Informatics and Information Technology, pp. 181–188. Institute of Informatics, PMF Skopje (2002)
4. Lim, C.C., Jin, J.S.: A Study on Applying Software Security to Information Systems: E-Learning Portals. IJCSNS International Journal of Computer Science and Network Security 6(3B), 161–166 (2006)
5. Rocklin, T.: Computers and testing. The National Teaching and Learning Forum 8(5), 1–4 (1999)
6. Advanced distributed learning, `http://www.adlnet.gov/scorm/index.aspx`
7. Pain, D., Le Heron, J.: WebCT and Online Assessment: The best thing since SOAP? Educational Technology & Society 6(2), 62–71 (2003)

Semantic Supported Modeling and Orchestration of Logistic Integrated Processes, with Focus on Supply Chain: Framework Design

Roberto Pérez López de Castro[1], Dania Pérez Armayor[2], Jorge Marx Gómez[3], Inty Sáez Mosquera[1], and José Antonio Díaz Batista[2]

[1] Central University of Las Villas, Cuba
{robertop,intysaez}@uclv.edu.cu
[2] Polytechnic University of Habana (CUJAE), Cuba
dania@ind.cujae.edu.cu
diaztony@tesla.cujae.edu.cu
[3] Carl von Ossietzky University Oldenburg, Germany
jorge.marx.gomez@uni-oldenburg.de

Abstract. Full (complete) integration is not yet achieved in supply networks; this is a complex challenge because of the importance of integration for management in this environment. Reducing the gap between semantic, Business Process Modeling and Interoperability solutions; will significantly (dramatically) improve the information flow in the chain and its understanding. For this purpose we present a proposal for a framework design that combines the semantic supported modeling with the orchestration of integration processes in the approached context. This will translate in better decisions based on latest and best information. There is also presented a possible support decision model that could be used for the validation of this framework.

Keywords: Semantic Business Process Modeling, Interoperability, Heterogeneity, SOA, Integration Technology Evaluation, Compensatory Fuzzy Logic.

1 Introduction

Inter-enterprise coordination is a core issue in Supply Chain Management (SCM). The challenge is to allow that every supply network member makes decisions based on the latest and best information from everyone else [1]. Companies invariably need to electronically exchange information and integrating such information into each member information system. The traditional solution consists on providing interfaces that allows the access of providers and clients to the necessary data for management [2]. However, because of the large number of diverse information systems, the data format (syntax) of each exchange (message) and the meaning (semantics) of both of them usually differs from company to company, or sometimes even within the same company if more than one software product is used as information system [2-5]. This makes it very challenging to exchange information in an interoperable way. Interoperability in this context means “the ability of two or more systems or components to exchange information and to use the information that has been exchanged” [6]. So, users either have to agree on a common model to express the data to be exchanged or they have to individually translate the data received from the business partner to the data format they own.

Despite the fact that process modeling languages allow the combination of process definition (their structure) with Web services orchestration (as process execution structure), they are not able until now to define the integration mechanisms for heterogeneous data schemas. On the other hand, the solutions designed to achieve interoperability of the information systems don’t achieve a complete integration since they don’t include the integration at processes level. The previously expressed features shows exactly the gap that our research will try to progressively reduce with the merging of the semantic supported process modeling in the frame of logistical management with focus on Supply Chain (in the tactical-strategic level), with their orchestration (via the usage of a service oriented architecture); reducing the integration efforts starting from the reduction of heterogeneity and uncertainty in decisions making in the tactical-strategic level of Supply Chain Management.

In this paper we present a proposal for the framework design that combines the semantic supported modeling with the orchestration of integration processes in the approached context; including the process model in the schema mapping of the exchanged data; and there is also presented a possible support decision model that could be used for the validation of this framework.

If integration technology could be associated to the requirements of a supply network type using truth grades as the extent to which these technologies meet the given integration requirements; this result can be used as a support decision model for the selection of technologies when implementing a new integration solution or to evaluate the combination of technologies in existing solutions, reducing the uncertainty related to it. It also can help software vendors or developers to identify the impact that their solutions can have in supply chains or to whish chain type to direct their efforts exposing the integration requirements that they should meet.

2 The Need for Integration

Traditionally there have been several systems for several activities, each application specializing on some specific activity type, in order to reduce time and increase effectiveness in each specific area of work. These autonomous and dissimilar systems were not made to collaborate among them, holding related data without the proper global administration [2]. Also in different department the same task were made using different software. This isolated information systems have a very negative impact on business efficiency and their effectiveness [7]. The more complex business processes brought the need for a combined effort among areas to give a coordinate answer to the clients regardless of how many internal systems were necessary to consult. In this environment the integration of information systems becomes significant in order to guarantee continuous and harmonic information flow in the enterprise frame [2, 3].

The necessity of integration of the information systems in the enterprise pave the way for the development of integration solutions, among them the most involving, and therefore the most popular have been the Enterprise Resource Planning (ERP) systems [1-3, 5, 8, 9]. The ERP systems provide an integration framework by re-implementing disparate enterprise application systems on the basis of an integrated and consistent database and accommodate many features of an organization's business processes, but they are highly complex, integrated systems that require careful consideration before selection, implementation, and use [2, 9] given the tremendous implementation cost and the difficulties to adapt the enterprise to the system [9, 10] specially in Small and Medium-sized Enterprises [11].

On the other hand, the increasingly competitive and time-sensitive market set the need for managing all companies involved in the flow of the same kind of goods and services as a coordinated system, in order to decrease global cost and time to market, and increase efficiency and customer service level. New types of information systems were developed to meet these new necessities, as Customer Relationship Management (CRM) systems and Supply Chain Management (SCM) systems. Now, those and other systems coexist in the supply chains in order to satisfy different needs, but they face similar problems to the ones used inside the enterprise frame regarding to the lack of ability to collaborate and the danger of data inconsistencies [2, 5, 7], the difference is that the tremendous complexity and costs do not allow to solve the problem using a single integrated system [2].

Supply networks need to work as a synchronized entity in order to reduce buffered inventory, lead times and, therefore, be able to reduce total cost across the entire system and increase responsiveness in order to match supply and demand in the market in ever-shorter time-frames. That is a complex task involving a great challenge of coordination and information integration in order to allow visibility and reduce uncertainty in the demand through the complete network [12]. In order to accomplish this every member of the network must be connected to each other by means of shared information [12]. This situation makes the integration of information systems across the whole network a necessary element in order to provide a shared foundation of information. The information systems are responsible for the visibility and for accomplishing that they must be design in relation to common predetermined objectives for all members involved. "The use of these systems has the potential to convert supply chains into demand chains in the sense that the system can now respond to known demand rather than having to anticipate that demand through a forecast." [12]

Nevertheless, to enable these benefits represents a tremendous task that requires a high level of process alignment [12]. The problems that make this task so difficult find their source in technological, strategic and organizational reasons. On one hand, there are findings arguing that companies share information selectively, they are often willing to share some information to ensure the flow of materials through the network, but very few seem to be willing to provide online access to their ERP-systems or access to sensitive areas such as design or strategic decisions, due to potential loss of proprietary information and loss of control [8].

Also there are many technologies to connect information systems, but none of them have claim to overcome all the integration problems and is necessary the use of combinations that are different according to the integration requirements of the supply networks [5]. So, the problem is which combination fits better a given integration necessity.

3 Integration Challenges

The process for enabling the previous mentioned benefits of the supply chain integration represent a tremendous task which progress is measured in years and even decades [1] confronting a great number of challenges. Several supply network partners regret the lack of visibility, the latency and inaccuracy of information gathering, and the impact on collaboration and confident decision making, influencing that many of the exchange of information is done using phone, fax, email, spreadsheets, etc. [13]. As was mentioned before sometimes managers and employees do not want to share their data, as they fear they would lose control over their processes, and as result their role might be diminished or eliminated [7, 8]. In top of that lies the resistance to change, the existence of many disparate systems or legacy systems that cannot be changed just to be connected to an integration solution, the cost to develop, implement, operate and maintain these integration solutions, the internal politics and the extra difficulties caused by the redesign of business processes [3, 7].

There are other several problems that can be associated to strategic, organizational or technical issues. Some strategic problems are related to low customer satisfaction, decline in market share, increased pressure from competitors, the inability to respond to customer demands [7]; also with maintaining balance between network and firm allegiance, financing network infrastructure, assessing and sharing risk and liability in the network, lack of frameworks for cooperation (security standards, legal standards, etc.) and establishing control and decision making in the network [4]. In the organizational related problems can be mentioned the mismatches between organizational and technical architectures, the semantic resolution of process, the data and process ownership disputes and others that make harder the collaboration between employees and departments across the organisations further impacting the business manageability [4, 7]. And finally, among the technical problems can be mentioned: tools and data sources can't always be standardized across network, identifying standards for data exchange and tool interoperability, maintaining a reliable infrastructure for communication, semantic resolution of data, and other issues also related to the organisation's information technologies infrastructure. [3, 4, 7]

4 The Gap + Idea

The integration costs for enterprise applications cooperation are still extremely high, because of different business processes, data organization, application interfaces that need to be reconciled, typically with great manual (and therefore error prone) intervention [14].

Towards the ultimate goal of seamless interaction among networked programs and devices, industry has developed orchestration and process modelling languages such as XLANG, WSFL and BPEL4WS (more recently WS-BPEL since version 2.0). Unfortunately, these efforts leave us a long way from seamless interoperation. Researchers in the Semantic Web community have taken up this challenge proposing top-down approaches to achieve aspects of Web Service inter operation. Unfortunately, many of these efforts have been disconnected from emerging industry standards, particularly in process modelling [15].

Business Process Management (BPM) has gained significant attention by both research and industry, and a multiplicity of BPM tools are already available and in use. However, the degree of mechanization in BPM is still very limited, creating inertia in the necessary evolution and dynamics of business processes.

There are several developed and ongoing researches on the topic [16-20] with the intention to include semantics in the process. On one side there are approaches applying ontologies to describe enterprise models and business processes in general, to show the potential benefits of the application of ontologies for companies; following with the automation of the transition from business process models to monitored execution, and the analysis of what went wrong using the business vocabulary that could be delivered by ontologies; and the latest steps, with attempts to automate processes using SOA and semantic web services.

Following these guidelines business analyst can use well-known flowchart-like graphics to model a new business process on his computer. This terminology can be matched to concepts and relations from ontologies; this means the process elements as ontology entities are specified in a machine-readable manner.

From these outputs, the executable description of the process is deployed. Semantic business process configuration involves composition (implementation of the process using web services), then translation from a business process modelling ontology to a semantically enhanced business process execution language, which is then further serialized to an executable specification. Finally, the executable process model is ready to be deployed to a process engine for execution.

With the development of these approaches business managers and analysts are acquiring very powerful tools, they can model new business processes, search for existing process fragments, automatically fill in the missing elements in the process model, search for semantic web services that will deliver the functionality, compose business processes out of available web services and execute implemented business process models; but never looking at the inconsistencies regarding the heterogeneity of the data representation schemas, specially related to cross company interaction, as it is the case of integration processes in a supply chain.

One of the most difficult problems in any integration effort is the missing interoperability at the data level. Frequently, the same concepts are embedded in different data models and represented differently. When facilitating interoperability at the data level one faces the problem that different data models are used as the basis for business formats. For example relational databases are based on the relational model, while XML Schema is basically a hierarchical model. The mapping process is intended to work on a neutral representation, which abstracts from the specific syntax and data model of a particular business schema definition. Therefore, all incoming business schemata are expressed in a neutral format. Emerging messaging systems, such as ebXML, need to include structural support for these semantics [14].

There are also approaches to consider in this area [21, 22]; however these solutions don’t achieve a complete integration because they don’t include the process level. As expressed before, we can see now that neither approach covers the integration requirements for an integrated process management in the addressed environment.

5 Details of the Framework

In this section we present a proposal for the framework design that combines the semantic supported modeling with the orchestration of integration processes in the approached context; including the process model in the schema mapping of the exchanged data.

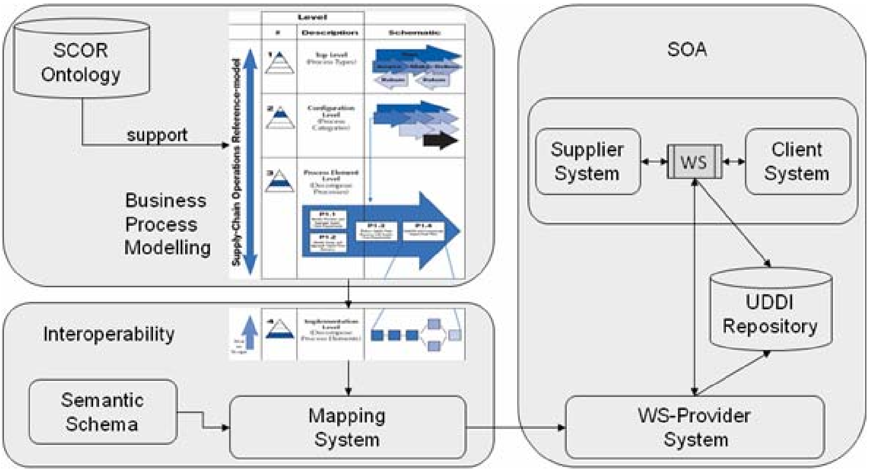


Fig. 1. Proposed Framework

The proposed framework combines semantic Business Process Modeling, with support from the SCOR (Supply-Chain Operations Reference-model) ontology; this will allow to model the logistic integration processes in the Supply Chain based on a recognized standard. The resulting model could be included then in the schema mapping process for the data elements associated to each process element, as the tree of elements should provide a hierarchy useful for the improvement of the procedure.

As result from the mapping system you get a transformation language (could be expressed in XSLT), it will be used to generate the web service which will be used as translator for the different schemas in the current information interchange. This way seamless interoperation at the process level is achieved including these translators in the executable process model of the workflow.

6 Validation

The technology development process could be extremely challenging, however there is also an open question regarding the evaluation of these technologies as exposed in several researches [5, 23, 24]. A way to reduce the uncertainty related to the selection of technologies when implementing a new integration solution, or to evaluate the combination of technologies in existing solutions, could be to associate these combinations to the requirements of a specific supply network type. Compensatory fuzzy logic could be applied to find the extent to which these technologies meet the given integration requirements.

Themistocleous, Irani and Love [5] proposed a framework for the evaluation of Enterprise Application Integration (EAI) technologies that can be used as reference to determine desirable permutations of integration technologies that can be used to unify applications. According to these authors findings the possible permutations of integration technologies must be based on: technologies functionality, integration requirements, and constrains of existing information systems infrastructures. The framework proposes a classification of the systems types that are generally integrated and the integration layers, using these two criteria, plus application elements, as technology evaluation criteria. As a result expose a qualitative classification of a set of EAI technologies. The analysis suffers from poor integration requirement description and recommends further generalization of the exposed findings.

This framework is being used as starting point for the development of a decision support model in order to reduce the uncertainty in the selection of integration technologies. The proposed model seeks to establish the foundations for a decision method that clarify the selection of "the best" technological combination given the supply chain requirements, assigning truth grades of integration technologies, or combinations of it, towards integration requirement of a given supply chain type using compensatory fuzzy logic.

Since the fit that a technology can reach depend on the characteristics of the supply chain, is recommendable to establish a classification of supply networks types (given that they are a main conditioning factor of the integration requirements), or at least a set of supply set of attributes that can be linked to specific requirements in a more detailed approach to a given logistic system.

Once the supply chain type was determined is possible establish the integration requirements that must be associated with the given supply chain, and the truth grades that express how related they are, as shown in Fig. 2, step 1.

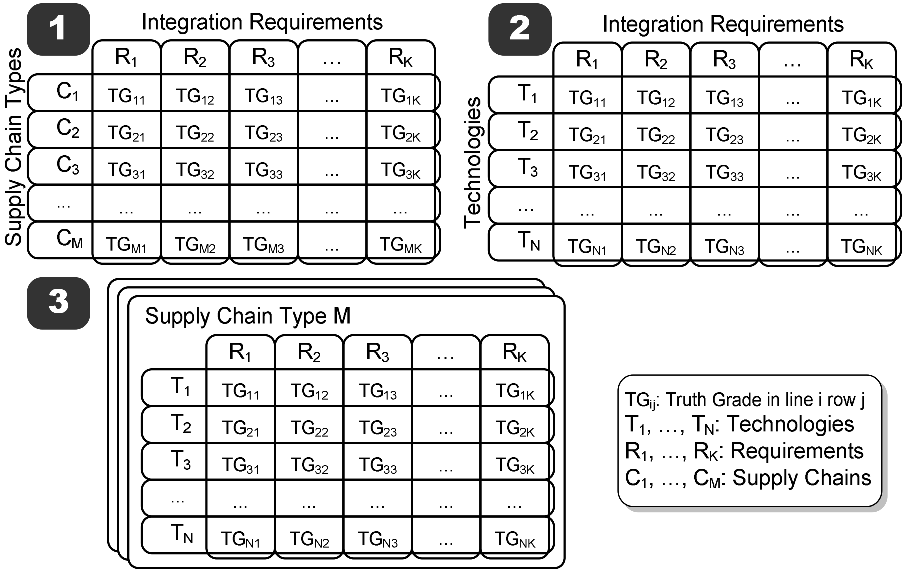
These requirements can help to distinguish the technologies or the combinations that must be evaluated for each chain type given their functionalities, for this will also be determine truth grades as a measure of the relation between requirements and technologies, as it appears in Fig. 2, step 2. To complete is also necessary establish a compatibility coefficient among technologies, otherwise the technology combinations can't be properly set up due the need to find the combinations of the most compatible technologies to each other.

At the end the idea is to determine the truth grades in a technology that fulfills the integration requirements of a certain chain, as the Fig. 2, step 3 illustrated.

The definition of supply chain types, integration requirements, technologies and technology combinations used in this analysis must be subjected to experts' criterion, using, for instance, Delphi Rounds[25], as a way to validate and strengthen partial findings of the ongoing investigation.

Also, in order to somehow quantify the pertinence that an integration requirement has with a supply chain and, equally, the fittingness that a technology has with an integration requirement the Compensatory Fuzzy Logic (CFL) will be used.

Compensatory Fuzzy Logic is a multivalent system that breaks with the traditional axioms of such systems to obtain a behavior semantically better than the classical systems. This approach abdicates to the compliance of the classic properties of the



	R_1	R_2	R_3	…	R_K
C_1	TG_{11}	TG_{12}	TG_{13}	...	TG_{1K}
C_2	TG_{21}	TG_{22}	TG_{23}	...	TG_{2K}
C_3	TG_{31}	TG_{32}	TG_{33}	...	TG_{3K}
...	...	...	...	...	...
C_M	TG_{M1}	TG_{M2}	TG_{M3}	...	TG_{MK}

	R_1	R_2	R_3	…	R_K
T_1	TG_{11}	TG_{12}	TG_{13}	...	TG_{1K}
T_2	TG_{21}	TG_{22}	TG_{23}	...	TG_{2K}
T_3	TG_{31}	TG_{32}	TG_{33}	...	TG_{3K}
...	...	...	...	...	...
T_N	TG_{N1}	TG_{N2}	TG_{N3}	...	TG_{NK}

	R_1	R_2	R_3	…	R_K
T_1	TG_{11}	TG_{12}	TG_{13}	...	TG_{1K}
T_2	TG_{21}	TG_{22}	TG_{23}	...	TG_{2K}
T_3	TG_{31}	TG_{32}	TG_{33}	...	TG_{3K}
...	...	...	...	...	...
T_N	TG_{N1}	TG_{N2}	TG_{N3}	...	TG_{NK}

TG_{ij}: Truth Grade in line i row j
T_1, …, T_N: Technologies
R_1, …, R_K: Requirements
C_1, …, C_M: Supply Chains

Fig. 2. Initial expected steps for a decision support model to assist in the selection of integration technologies for a supply network.

conjunction and the disjunction, contrasting the idea that the increase or decrease in values of either true-induced change of truth value of one of its components can be compensated with a corresponding decrease or increase in other. It is the first multivalent system distinguishes with the property to generalize the Bivalent Logic completely. Its capacity to formalize the reasoning makes it feasible to use for situations that require multi-criteria evaluations and verbal descriptions of knowledge, which often are described in ambiguous form; therefore, it is an opportunity to use language to build semantic models that facilitate the evaluation, the decision making and the discovery of knowledge. [26, 27]

7 Conclusions and Future Work

Integration solutions are critical component of today's enterprise strategies, but it's a long way between the high-level vision of the integrated supply network and the basic reality in the development and implementation of these solutions.

A contribution to solve this problem is presented by means of a proposal for a framework design that combines the semantic supported modeling with the orchestration of integration processes in the approached context.

The selection of technologies when implementing a new integration solution or to evaluate the combination of technologies in existing solutions is a strongly discussed issue nowadays that could be diminished using a decision model that associate integration technology combinations to the requirements of a supply network. Compensatory Fuzzy Logic is a superior approach that will be used to achieve this purpose.

As future work is projected the implementation of a prototype for the proposed framework, and is also planned to define the theoretical bases for the formulation of the decision model anticipated as a way for the validation of this framework.

References

1. Davenport, T.H., Brooks, J.D.: Enterprise systems and the supply chain. Journal of Enterprise Information Management 17(1), 8–19 (2004)
2. Weske, M.: Business Process Management. Concepts, Languages, Architectures. Springer, Heidelberg (2007)
3. Hohpe, G., Woolf, B.: Enterprise Integration Patterns: Designing, Building, and Deploying Messaging Solutions. Addison-Wesley, Reading (2004)
4. Glazner, C.G.: Enterprise integration strategies across virtual extended enterprise networks: a case study of the F-35 Joint Strike Fighter Program enterprise, in Massachusetts Institute of Technology. Technology and Policy Program, Massachusetts Institute of Technology: United States of America (2006)
5. Themistocleous, M., Irani, Z., Love, P.E.D.: Evaluating the integration of supply chain information systems: a case study. European Journal of Operational Research 159(2), 393–405 (2004)
6. IEEE Standards, IEEE Standard Computer Dictionary in A Compilation of IEEE Standard Computer Glossaries, Institute of Electrical and Electronics Engineers, IEEE (1990)
7. Woznica, J., Healy, K.: The level of information systems integration in SMEs in Irish manufacturing sector. Journal of Small Business and Enterprise Development 16(1), 115–130 (2009)
8. Bagchi, P.K., et al.: Supply chain integration: a European survey. The International Journal of Logistics Management 16(2), 275–294 (2005)
9. Adam, F., Sammon, D. (eds.): The Enterprise Resource Planning Decade: Lessons Learned and Issues for the Future. Idea Group Publishing (2004)
10. Wallace, T.F., Kremzar, M.H.: ERP: Making It Happen. The Implementers' Guide to Success with Enterprise Resource Planning. John Wiley & Sons Inc., New Jersey (2001)
11. Brehm, N.: Föderierte ERP-Systeme auf Basis von Web Services. Dissertation. Shaker Verlag, Aachen (2009)
12. Christopher, M.: Logistics & Supply Chain Management: creating value-adding networks, 3rd edn. Financial Times Series, p. 320. Prentice Hall, Harlow (2005)
13. Rollings, S.: Does the World Need Another Supply Chain Application? Outsourced Logistics, 30–33 (2008)
14. Beneventano, D., et al.: Ontology-driven Semantic Mapping. In: Conference Proceedings of I-ESA 2008 (2008)
15. Mandell, D.J., McIlraith, S.A.: Adapting BPEL4WS for the Semantic Web: The Bottom-Up Approach to Web Service Interoperation. In: Fensel, D., Sycara, K., Mylopoulos, J. (eds.) ISWC 2003. LNCS, vol. 2870, pp. 227–241. Springer, Heidelberg (2003)
16. Fox, M.: The TOVE Project: A common-sense model of the enterprise. In: Belli, F., Radermacher, F.J. (eds.) IEA/AIE 1992. LNCS, vol. 604, pp. 25–34. Springer, Heidelberg (1992)
17. Lampe, J.C.: Discussion of an ontological analysis of the economic primitives of the extended-REA enterprise information architecture. International Journal of Accounting Information Systems 3(1) (2002)
18. e3-value, `http://www.e3value.com`

19. SUPER Project, `http://www.ip-super.org`
20. Uschold, M., et al.: The Enterprise Ontology. The Knowledge Engineer Review 13(1) (1998)
21. Heflin, J., Hendler, J., Luke, S.: SHOE: A Blueprint for the Semantic Web. In: Fensel, D., et al. (eds.) Spinning the Semantic Web. MIT Press, Cambridge (2003)
22. STASIS Project, `http://www.stasis-project.net/`
23. Sharif, A.M., Irani, Z.: Exploring Fuzzy Cognitive Mapping for IS Evaluation. European Journal of Operational Research 173(3), 1175–1187 (2006)
24. Khoumbati, K., Themistocleous, M.: Application of fuzzy simulation for the evaluation of enterprise application integration in healthcare organisations. Transforming Government: People, Process and Policy 1(3), 230–241 (2007)
25. Linstone, H.A., Turoff, M. (eds.): The Delphi Method:Techniques and Applications. New Jersey Institute of Technology, New Jersey (2002)
26. Espín Andrade, R.A., Fernández González, E.: La lógica difusa compensatoria: una plataforma para el razonamiento y la representación del conocimiento en un ambiente de decisión multicriterio. In: Leyva López, J.C. (ed.) Análisis Multicriterio para la Toma de Decisiones: Métodos y Aplicaciones, editorial Plaza y Valdes; editorial Universidad de Occidente (2009)
27. Martínez Alonso, M., et al.: Experiencias en el descubrimiento de conocimientos a partir de la obtención de predicados en lógica difusa compensatoria. In: Segundo Taller de Descubrimiento de Conocimiento, Gestión del Conocimiento y Toma de Decisiones. Ciudad de Panamá, Panamá (2009)

Web Service Validation within Semantic SOA-Based Model

Tariq Mahmoud, Timo von der Dovenmühle, and Jorge Marx Gómez

Department of Business Informatics I, Carl von Ossietzky University Oldenburg,
Ammerländer Heerstr. 114-118, 26129 Oldenburg, Germany
mahmoud@wi-ol.de, Timo.von.der.Dovenmuehle@Uni-Oldenburg.de,
marx-gomez@wi-ol.de

Abstract. Nowadays, it is getting more and more difficult discovering the entire Web resources in order to find the correct information needed by user. The correctness of the retrieved results in most cases is not meeting the expected needs because of the weak definition requests provided by the user and this came from the unstructured nature of the existing information resources.

The main problems SOA solutions faces are the lack of automation at both discovery and invocation phases among these services and the queries provided by the consumers. Our proposed solution will overcome the poorness of traditional SOA solutions and the complexity of the semantic ones by developing a consistent framework that makes the data understandable for both humans and machines. In addition it will provide Web Service validation and a methodology for dynamic composition of Web Services.

Keywords: SOA, Semantic SOA, Web Service Validation, Ontology, Ontology Model.

1 Introduction and Problem Definition

In these times, it becomes very hard for anybody in the digital world to search and find suitable services fit into his/her needs, since there is a huge amount of data on the Web caused by the enormous increasing of the Web documents, Web users and Web Services provided in this digital community. And the most difficulties Web Services have to overcome, in the attempt to use the contents of the World Wide Web [1], are heterogeneity which is caused by the nature of the Web itself side by side with the issues regarding information sharing, finding, extraction, interpreting, maintaining and representation [2].

One powerful concept involved in the context of information sharing environments is the Service Oriented Architecture (SOA) concept and the usage of Web Services (WS) in order to enable machine to machine interactions over the Web.

Because the information type is the basic principle which a system has to deal with, it is important to understand the different types of information that this system has to manage. And there are three different types of information can be mentioned in this context: internal, semi-internal and external information [3].

The internal information is based on the data produced inside an organization. It has to be audited, documented, archived and made available to any interested potential party. The domain model here is also internal and this internal information is correct within the organization view. And as a result the error rate is indeterminable because the internal validation of this information will not point out mistakes as this information is "correct".

The semi-internal information is produced by external systems to be dedicated for an organization. In addition to the issues related to internal information, the semi-internal information has to be understandable. At some points, the semi-internal information does not meet the used domain model specification and this means that the semi-internal information has to be validated.

The external information is produced for different goals outside an organization. As a result, first of all the context of the information has to be defined and retrieved before it can be validated. The challenge of validating external information is the used domain model specifications and sometimes these specifications look similar to the targeted ones. However, not always the information representation methods are similar and that will be difficult to recognize so validating external information is important if it has to be used within an organization.

Nowadays, information systems are analyzing syntactic data and the meaning of information is hidden in documents and other storage systems so it is nearly impossible to extract the meaning of information via algorithms, because of the unstructured nature of the data. A non-technical problem here is the absence of research knowledge in this area and most people do research in a nonprofessional manner without using the meaning of information. And the challenges will be in finding the right resource for a consumer and to face the problem of lacking the relevant Meta data related to the information he seeks. Last and not least there is a great risk in getting misused information that doesn't meet the consumer needs.

Semantic Web [4] is an evolving extension of the existing Web in a way that the semantics of information and services on the Web is defined, making it possible for the Web to understand and satisfy the requests of people and machines to use the Web content.

However, from one side the new emerging technologies in the world of Semantic Web makes the Semantic SOA techniques seem to be inaccurate to be used in terms of semanticizing the capabilities of Web Services and the requests of WS consumers because of the blurred representation of the involved ontology. And from the other side, traditional SOA-based solutions lack semantic documentation of WS interfaces [5], and that will return misused information to the consumer as we had mentioned above.

Based on that, our proposed light weight semantic SOA-based solution tries to make the data understandable for both humans and machines and it will have the responsibility of splitting the semantic annotation from the core services in a way that both normal and Semantic Web Services (SWS) [6] can be validated and used [7]. This model will also provide a second level of WS classification by grouping Web Services in categories based on the area of interest named "WS clouds" which will entail their concepts from a predefined ontology, and this will be explained in details later in this paper.

The rest of our paper is structured as follows: in section two we provide a short overview on the methodology and the architecture of the proposed semantic SOA-based model that we are developing as a work on progress. And in section three, we intend to analyze the validation of Web Services within the semantic SOA-based model by illustrating service alliance and explaining the design aspects. Finally we will give a brief conclusion with a snapshot to our future work.

2 Semantic SOA-Based Model

The main idea behind this model is to have an ontology that has the role of dealing with Semantic Web Services as well as representing the whole concepts of a Web Service; it has also category type and generic operations divided to syntactic, semantic, behavioral and qualitative operations. Starting from this point, new categorizing level has to be done by the help of the cloud provider (see next paragraph) who will have the responsibility of grouping WSs in clouds. The classifying will be based on the domain of interest, each of these clouds is itself a WS and the cloud provider will advertise them in the WS directory.

At this point the WS provider system will search in the directory for its cloud based on its business domain and interests and will register itself in one of them to later on entail the required concepts out of that cloud and implement its own WSs to be registered also in it, and if there is no cloud matches with its purposes, the service provider will ask the cloud provider to create a new one and advertise it in the directory as well, and here the concepts will be inherit from the metadata ontology.

The process of adding semantics to WS request and capability will take benefit from the semantic mediator-based system functionality by submitting semantic goal (highly defined WS requests) to be fulfilled with SWS capability. The mediation issues will be also applied depending on the variety of mediators provided by the mediator-based system as well.

The new architecture is depicted in Fig 1 and it shows the following components:

- User System: Is the subsystem which implements functions that will be used in the end users interfaces. This subsystem is able to generate user screens at runtime.
- Workflow System: Deals with business processes described in an appropriate XML-based workflow language. A workflow in this system is a plan of sequentially or in parallel chained functions as activities in order to create or utilize business processes. Workflows implicitly contain the business logic of the overall system because it is the main unit in it and all the activities will be distributed to the other subsystems starting from here.
- Web Service Consumer System: Contains XML schema definitions and functions needed for the processes of WS discovery and invocation provided by different service providers.

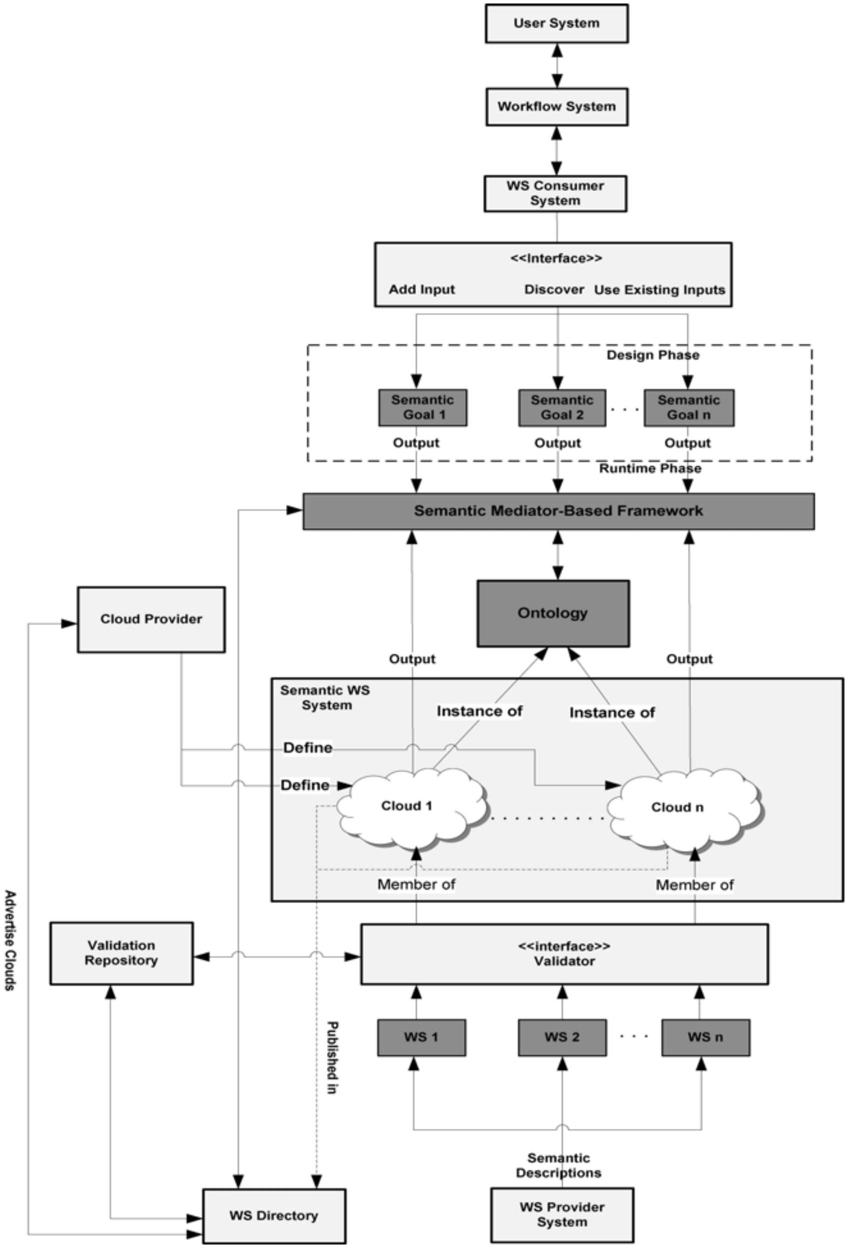


Fig. 1. Semantic SOA-based Model for Business Applications

- Web Service Provider System: Contains functions required for providing Web Services and it is dealing with HTTP incoming and outgoing user's requests, and has a connection to Web Service directory via the validator interface in order to allow the publication of Web Services after performing the validation process for them.

- Web Service Directory: Its interface has the responsibility of the publication and searching for Web Services based on a semantic goal provided by the consuming system and the Universal Description Discovery and Integration (UDDI) standard [8].
- Semantic Web Service System: The main two entities in this component are the ontology and the cloud of Web Services. Ontology here will be a metadata ontology that will serve as a base to provide concepts to help the service and cloud providers in the process of creating the clouds and the Web Services within the clouds by enabling them to derive the descriptions of the concepts in the creation process. While the clouds here are defined by the cloud provider as instances of the metadata ontology.
    - The cloud is a service itself, and it can be created, advertised, discovered and invoked in the traditional way. In addition, they are published in the WS directory in order to let the service providers discover them to later on register their services as members within them. Each cloud has a category type and generic operations.
- Cloud Provider: Will define the clouds as instances of the metadata ontology by assigning values to its concepts. The cloud providers can be normal service providers or businesses that share common point of interest (in most cases, profit organizations).
- Semantic Mediator-based System: Has the responsibility of solving the heterogeneity issues that might occur between the semantic goals provided by WS consumer system and the semantic descriptions of Web Services given by the WS provider system by performing the matchmaking and filtering results processes. The mediator-based system may be embedded in a middleware as depicted in the architecture, or it can be an external WS which accomplishes the mediating scenarios at run-time in a way that lessens the load of mediation process. This loose coupling promotes reusability and facilitates dynamic partner binding, especially at runtime. And all of these issues are to be considered in our future work.
- Validator: It is an interface which has the responsibility of tunnelling the communication between the WS capabilities and concepts derived from the clouds in the semantic WS system and also monitoring the non-functional properties of WS.
- Validation Repository: It has the functionality of calculating the values of non-functional properties in order to forward it to the validator interface to make the proper mapping decisions and this repository also contains the concepts relations that are entailed from the existing domain ontologies outside this model. It has also the task of publishing the WS provider's WSs in the directory since it has a link to the WS directory to register its WSs that had passed the validation stage successfully.

The main outcomes from using this model over the existing approaches appears in the high ability of reusing the functionality which means that each component in the architecture can be considered as a standalone input of the system. Another added value can be generating dynamic workflows in the process of composing a new service that is not existed in the system and this composition will be dynamic because the Web Services in this architecture have the same root metadata ontology so it will be easier to compose services based on the consumes semantic goals. Moreover, categorizing the Web Services in clouds will provide a second level of classification together with the traditional one in the directories and this will make the WS searching mechanism more powerful and will improve the response time of the discovery phase activities and the overall performance. Finally, there will be an advertisement for new Web Services in the case of desired functionality absence by forwarding the concepts to the cloud provider in order to create new cloud that motivates the WS providers to create WS having similar functionality.

3 Web Service Validation within Semantic SOA-Based Model

3.1 Web Service Compliance

One goal of the semantic SOA-based model is to annotate information with semantic Meta data. In contrast to other concepts, the annotation [9] here is not a part of the adopted Web Service itself. There is an isolated service, which is used to perform this annotation task. It is a precondition to validate the adopted WS against the external domain ontologies. The reason behind that exists in the issue that the user of an ontology expects that the service will work as the ontology describes. And the only way to insure the behavior is to make service validation.

We can classify the compliance of Web Services in four main different types:

- Exact: WS is able to comply with the requirements (WS properties).
- Over-Exact: WS has higher compliance with the requirements more than expected.
- Partial: WS is able to comply with the requirements fractionally.
- Failure: WS is not able to comply with the requirements.

An exact compliance is the ideal situation. In this case, a WS fulfills the expectations. An over-exact compliance is happening, when the WS provides a higher level of quality more than expected by the consumer. The partial compliance of a WS is happened when the requirements can be divided into logical parts. If the WS complies with a part, then it can be used to handle the information in collaboration with another Web Services. Failure compliance is happened when the WS does not comply with the request or parts of that request.

To clarify the abovementioned compliance types we can give the following example: suppose a data set about a customer that has to be validated and stored and it contains his first and last names, email and birth date. The validation of the email address and the birth date can be done partially and independently from storing the customer's information.

$$O_{request} := \{lastName, firstName, birthDate\} \tag{1}$$

$$O_{divided} := \{d_1(lastName, firstName); d_2(email); d_3(birthDate)\} \tag{2}$$

Web Services WS_1 and WS_2 are partially complaint to the request where WS_1 complies with the customer's email address and WS_2 complies with his birth date. They are used to handle the validation. WS_3 is an exact match because it complies with all of the customer's information and it will be used to store this information.

$$Mapping_{partial}\{d_2 \Rightarrow WS_1; d_3 \Rightarrow WS_2\} \tag{3}$$

$$Mapping_{exact}\{d_1, d_2(WS_1), d_3(WS_2) \Rightarrow WS_3\} \tag{4}$$

From a provider's perspective, this provides a possibility to integrate low capacity Web Services into ambitious requests by making crossing among the possible requests. But the challenge is to analyze the structure of the information by machines not by humans. To do so, the person within the ontology has to be broken down into properties. Typically, literals are used to store information like the name of a person and in this way, a first name can be everything described as an array of characters. However, in the real world this is not totally true because there are small quantities of character combinations that are valid to be first names. Considering this fact, the correct data type for the property first name would be an enumeration or an object that contains a table of valid entries. An email address is not a random literal too and the previous example showed that the design of the ontology has to start at the level of the properties not the level of entities.

Another challenge is the reflection of an ontology to an information system. Even at the level of primitive data types, there are many possible incompatibilities. The size of an xsd:string for example is limited by the storage file system [10]. It is not a big deal to create an xsd:string with the size of 3 Gbytes but it is not possible to load this string into a Java-based software because the Java:string length is limited to $(2^{31} - 1)$ bytes. If the data has to be stored in a database, there is another problem: which size is the right one for a text field? Another issues can be the handling of numbers, so if an xsd:PositiveInteger has to be handled, from one side the same problems will be appeared and from the other side there is no PositiveInteger as a data type in a lot of software platforms, and this will give the possibility to deal with such invalid values from the ontology's perspective.

3.2 Design Aspects

As an answer to abovementioned problems we can define ontologies at the lowest possible level and generalize more complex ontologies. By doing so, it is important to think about two important points: Firstly, the ontology has to be implementable to work in an information system and secondly, the ontology definition has to be done by domain experts who define all the concepts that will are composing such ontology. The implementation needs to have positive and negative validation tests in order to prove the correctness of valid targeting range. Where the positive validation is occurred within the range of the targeting values and the negative validation occurs within the input values that are invalid and outside the range of targeting values.

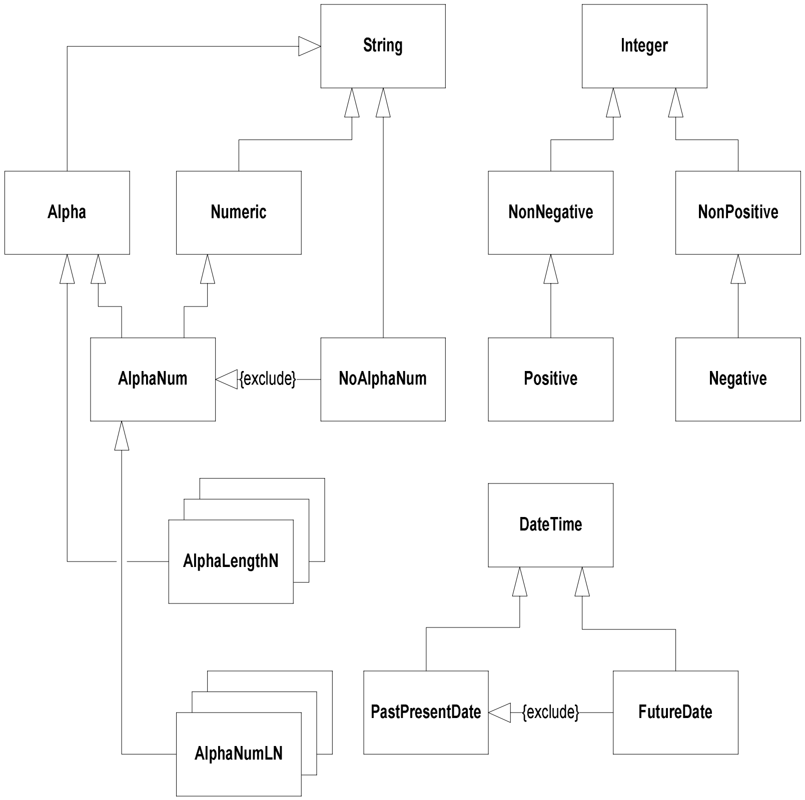


Fig. 2. Inheritance of Type Properties

The question is: why should we do that? Typically, in an information system there are many layers where data is validated. The first layer is the user interface, where the user input is validated. However, even the user acts as a validator, because he knows the domain and can filter evidently wrong data. At the next layer, the business logic deals with the data and check the plausibility and this might be repeated at the lower layers consequently.

This approach is correct, if we are working with an isolated system. But in SOA solutions, there is a good chance that different applications are using the same service, if these applications share data then the following problem might appear: an application assumes that the stored data is correct, and if the other application validates the data at a lower quality level, there is a risk to load invalid data. The only way to prevent this is to validate the data at the target service. This means, it is necessary that an object does not have only properties, attributes and methods to be existed rather it must be able to validate its own status!

4 Conclusion and Outlook

In this paper, the main focus was to introduce an ontological SOA-based architecture that deals with the Semantic Web Services, in order to be applied later on to one of the business solutions; also we had presented Web Service validation within this model by explaining the different Web Service compliance types that are exact, over-exact, partial and failure.

One of the main purposes of this architecture is to group the Web Services based on the actual domain that they are related to (the area of interest). In addition, we have the vision of creating general ontology to compose ERP-related Semantic Web Services out of it in order to have a general knowledge base shared between several Web Service providers based completely on ERP concepts.

Moreover, we had presented in this paper an ontology model that describes how we can get more complex ontologies out of the lower-level ones.

The upcoming implementations will include the process of implementing ontology to create Web Services clouds, grouping Web Services in these clouds, defining semantic goals, performing the matchmaking process between goals and Web Services, creating static and dynamic workflows and implementing them using workflow engines.

There is now an ongoing research for testing the quality of Web Services that are used in this architecture. Trust and security issues also will be part of the future work.

A concrete methodology for performing the dynamic Web Service composition will be defined in the future work.

Finally, prototype implementations of the future will show the practicability of those concepts.

References

1. Berners-Lee, T., Calliau, R.: WorldWideWeb: Proposal for a HyperText Project (1990)
2. Mahmoud, T., Marx Gómez, J.: Towards Process Mediation in Semantic Service Oriented Architecture. In: Handbook of Research on Social Dimensions of Semantic Technologies and Web Services, pp. 780–802. IGI Global (2009)
3. Sheth, A.: Enterprise Applications Of Semantic Web: The Sweet Spot Of Risk And Compliance. In: IASW 2005: International Conference on Industrial Applications of Semantic Web, Jyväskyla, Finland (2005)
4. Berners-Lee, T., Hendler, J., Lassila, O.: The Semantic Web. Scientific American Magazine, May 17 (2001)
5. Mahmoud, T., Marx Gómez, J.: Semantic Web Services Process Mediation Using WSMX Concepts. In: Proceedings 20th International Conference on Systems Research, Informatics and Cybernetics (InterSymp-2008) Baden-Baden, Germany (2008)
6. Studer, R., Grimm, S., Abecker, A.: Semantic Web Services: Concept, Technologies and Applications, pp. 287–309. Springer, Heidelberg (2007)
7. Maximilien, E., Munindar, P.: Towards Autonomic Web Services Trust and Selection, pp. 212–221. ACM, New York (2004)
8. Clement, L., Hately, A., von Riegen, C., Rogers, T.: UDDI version 3.0.2. UDDI Spec Technical Committee Draft (2004), http://uddi.org/pubs/uddi_v3.htm
9. Handschuh, S., Staab, S.: Annotation for the Semantic Web (2003)
10. Peterson, D., Gao, S., Malhotra, A., Sperberg-McQueen, C.M., Thompson, H.: W3C XML Schema Definition Language (XSD) 1.1 Part 2: Datatypes. W3C, http://www.w3.org/TR/xmlschema11-2/ (30.04 2009)

Auto-production 3D Graphics Content for Mobile Communication

Blagica Jovanova, Marius Preda, and Françoise Preteux

Télécom & Management SudParis (ex INT)
9 rue Charles Fourier, 91011 Évry Cedex, France
{Blagica.Jovanova,Marius.Preda,Francoise.Preteux}@it-sudparis.eu

Abstract. In this paper we present a new mobile service that enable auto-production of 3D graphics content for mobile platform. The general benefit of this service is that it enriches the content of the communication, containing a video stream of an animated avatars produced by dedicated servers controlled by the mobile end-user. The approach consists in selecting an avatar, download it on the mobile phone as a 3D object and composing the message by playing with avatar instead of typing text. Message can be enriched more by adding background to the scene and subtitles for animation sub-sequences. The tests performed show the feasibility of the proposed solution in terms of transmission cost.

Keywords: mobile communication, 3D graphics, auto-production, avatars, computer animation.

1 Introduction

Nowadays, SMS/MMS communication is a commercial success, even their relative poor communication quality, both in terms of information (re)presentation – textual data, and input interface – mobile keyboard. SMS/MMS is not just a way of simple communication, but it is also a way of expression. Some new dictionaries of symbols containing different emotions try to enrich the SMS quality, but they are still away from what the user wants to express. Techniques known as 'auto-production'[1] enables a user to create own content on mobile phone, i.e. to combine ring tones, images or videos and to add them in the message. Until now auto-production using 3D graphics content is not known.

In this paper we propose a solution to "pseudo-create" 3D graphics content on the mobile phone. It consists in a client-server infrastructure for creating the content, converting and transmitting it to a mobile-phone. The usage of the processing server is required since mobile platforms are generally too weak to perform computations needed for high quality 3D graphics. In short, two main components exists: a light JAVA application installed on the mobile phone that allows to select an avatar, pre-recorded animations and environments or to create new animations, and a server

[1] http://www.trakax.com/software/mobile/

database containing all the 3D assets in both high and low resolution. Once the sender finalizes editing the scene (as a combination of avatar ID, animations IDs and environments IDs) the parameters are transmitted to the server, the latter producing a video stream by rendering high resolution version of the scene. Finally the video is transmitted to the receiver's mobile phone. By using a web interface anyone can upload, visualize and manage avatars, animation and backgrounds that can be used afterwards by the service. The 3D files stored on the server are represented in M3G for low-resolution and MPEG-4 3D for high resolution.

The proposed prototype was evaluated objectively (measuring data transmission, message production time …) proving the pertinence of using avatars when the message to be sent has emotional insight.

The remaining of the paper is organized as follows. Section 2 gives a brief description of avatars technologies. Section 3 introduces software architecture. Section 4 presents the quantitative and qualitative results. We conclude the paper in Section 5.

2 Avatar Modeling and Animation

It become more and more evident that Internet will migrate from a repository of information to a dynamic and lively place where people communicate with each other and jointly interact with the content, where time, presence and events become important. Virtual Worlds (VWs), and especially 3D Graphics VWs, are now well-known applications. Initially conceived for social purposes as a communication support (i.e. chatting) and offering awareness on the presence and mood of the interlocutors, 3D VWs are now reaching a milestone: the technology for creating, representing and visualizing 3D content becomes available and accessible on different platforms including mobile devices. In this context, having a visual representation of users by means of avatars holding personal information, history, personality, skills, etc. becomes a necessity. In the mobile context, the feasibility of VWs is still to be proved. Several attempts were reported for using mobile phones as end-user terminals for visualizing 3D graphics worlds mainly based on content adaptation [20, 1] or remote rendering [27]. The real-time requirement of such applications imposes additional challenges to the terminal and the network. A less constrained application from the perspective of network capabilities is the so-called avatar personalization as extension of the Tamagotchi [25] concept. Modern phones have enough processing power (some of them coming with 3D hardware acceleration) to perform real-time animation of simplified avatars. The most difficult operations when considering avatars are their creation and, for the specific case of mobiles, their simplification while maintaining accurate representation. The day by day experience in observing human beings makes the human brain a powerful system able to observe without effort any non-natural effect of avatar modeling or animation. In the following we describe several approaches of avatar modeling and animation.

Two different courses can be chosen from in order to build a virtual character according to the researched appearance of the virtual character (cartoon-like or realistic), and depending on the technology available to the designer.

On one hand, the designer can build interactively the model's anatomical segments and set up the model hierarchy. In addition to creating the geometry and texturing it (both representing the avatar appearance) it is also needed to set up the skeleton and link it to the mesh. Several authoring tools and geometry generating mechanisms make it possible to model a virtual character [1, 15]. While the technique gives excellent results, its main drawback is that the result is strongly dependent on the designer's artistic skills and experience. In addition, this procedure is tedious and time-consuming.

On other hand, a faster and proven method is the use of 3D scanners. Contrary to computer aided design, the aim of 3D scanning is to create an electronic representation of an existing object, capturing its shape, color, reflectance or other visual properties. In its principle, 3D scanning is similar to a number of other important technologies (like photocopying and video) that quickly, accurately and cheaply record useful aspects of physical reality. A recent trend supported by the development of vision systems consists in capturing real persons by using one or several cameras and reconstruct or modify an existing template by using real measurements. In the case of monocular images such as in [10] the geometry obtained is mapped on a previously created model, providing a cheap and useful approach for automatic modeling. By using stereo or general multi-view systems, the 3D geometry may be recovered more accurately. Introducing anthropometry (studying and collecting human variability in faces and bodies) in computer graphics [3] made possible the creation of a parametric model defined as a linear combination of templates. The basis is extracted from large databases including human measurements such as NASA Man-Systems Integration Standard [18] and the Anthropometry Source Book [19], and several methods in exploiting it are provided in [22, 7].

In former published papers, avatars motion models were based on simplified human skeleton with joints [26]. More realistic deformation was achieved by adding new layers in addition to the skeleton, namely muscle, fatty tissue, skin and clothing [28, 6, 21, 23]. In our application we used artist-made characters because we observed that a caricature style is preferred against a realistic one in order to emphasize emotional content.

Once the virtual character has been created, one should be able to change its postures in order to obtain the desired animation effect. Animating a seamless character consists in applying deformations at the skin level. The major 3D mesh deformation approaches can be classified into the following 5 categories: lattice [14], cluster [14], spline [4], morphing [5] and skeleton [13]. The first four categories are used in animating specific objects as eyes (lattice), and face expressions (morphing), and are more or less supported by the main animation software packages. The last category, more and more encountered in virtual character animation systems, introduces the concept of skeleton. To design the virtual character skeleton, an initialization stage is necessary: the designer has to specify the influence region of each bone of the skeleton as well as a measure of influence. This stage is mostly interactive and recursively repeated until the desired animation effects are reached. When the skeleton moves, the new position of the vertex is calculated by multiplying the old position with the weights and matrices of the parent bones. While simple and easy to implement, the technique has some limitations, especially when animating soft body (the elbow problem).

Once the avatar appearance and skeleton set-up, to animate the avatar consists in defining the geometric transforms of bones. Current authoring tools dispose of a rich set of techniques for specifying joint angles. A classical solution is to directly control the relative geometric transformation of each bone of the skeleton, approach called Forward Kinematics (FK) [29]. An alternative approach is to fix the location in the world coordinates for a specific level of the skeleton, so-called end-effector (e.g. the hand for a human avatar), and to adjust accordingly the geometric transformation of its parents in the skeleton, method called Inverse Kinematics (IK). To achieve more realistic movements, motion capture technique [16] may be used. It consists in tracking and recording the position (and the orientation) of a set of markers placed on the surface of a real object. Usually, the markers are positioned at the joints. The markers' positions, expressed in the world coordinate system, are then converted into a set of geometric transformations for each joint [3, 11, 17].

3 Software Architecture

The proposed system architecture is illustrated in Fig. 1. and consists in two main modules, connected at the level of data repository.

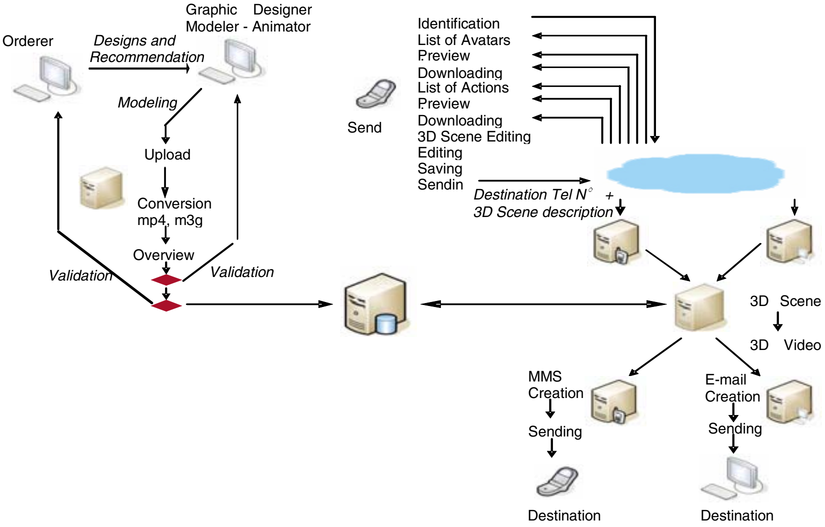


Fig. 1. System architecture for content creators (left) and end-user application (right)

The first is reserved to content creators (left), and allows uploading to the central repository and processing (converting) new content. The second (right) consists in an application server and a mobile client component which enables the end-user to create the animated messages.

3.1 Content Creator Module

The web-interface of the content creation module is illustrated in Fig. 2. It serves to upload 3D objects and animations as well as backgrounds. Currently the user can upload 3D data that is compliant with 3DSMax and it is internally converted in MPEG-4 3D graphics format.

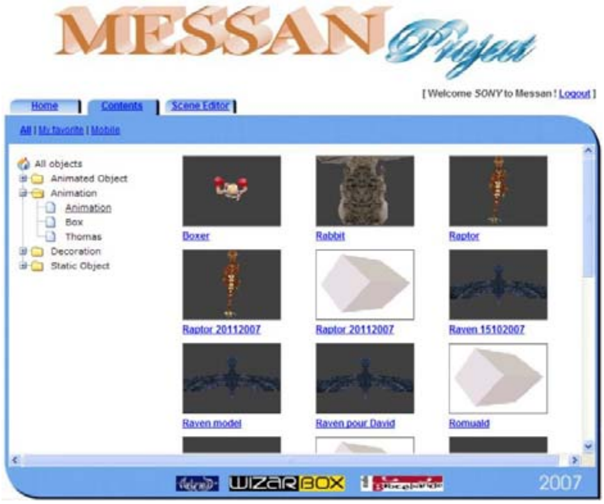


Fig. 2. The web interface for content creators

Both high definition (HD) and low definition (LD) version of an object should be uploaded. This is necessary because the LD version is transmitted to the mobile device for optimized visualization, and the HD version is used on server side for creating the final scene. Also video and image files may be uploaded and these are used for preview on the mobile device.

3.2 End-User Module

The end-user module has two components: a client for composing the animated message and an application server that produce the video based on instructions received from the mobile.

3.2.1 Mobile Client

This component is installed on the mobile phone and allows the end-user to compose the personal animated message. The scenario of creating the scene is shown in the diagram illustrated in Fig. 3. First, the application exposes the avatars available as a list of icons. User can select one and download it. He can start creating messages in different scenarios:

- he can asks for list of animations, select some of them and add in his scene;
- he can move the avatar, rotate it and record his own animation;
- he can ask for list of backgrounds and select some of them;

- on each animation he can add text that will be displayed like a subtitle on the animation
- he can add simple standard SMS text that will appear like a usual SMS at the end of the animation

Several animations can be appended and mixing between pre-created animations with the ones created in real-time is possible. After the user finishes editing, all parameters for composing the scene are sent to the server.

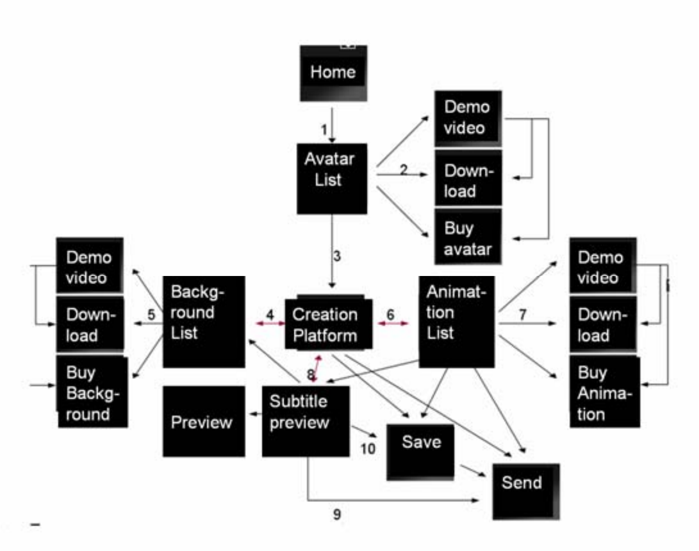


Fig. 3. End-user protocol for creating the animated message

3.2.2 Application Server

Once receiving all the composition parameters, the server proceeds with the creation of the 3D scene. Fig. 4. shows the main components implemented on the server.

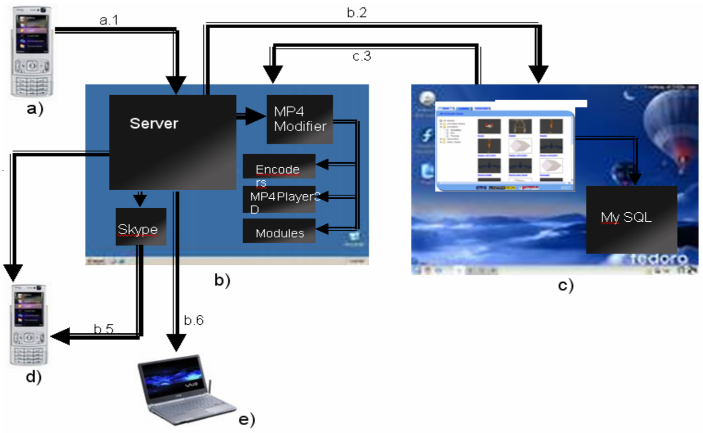


Fig. 4. Server protocol for creating the animated message

From mobile phone (a) only small amount of information is sent to the server (interface a.1): the type of request, the avatar's ID, background ID, animation IDs, the textual message, timestamps, destination number, etc. The server makes requests (b.2) to the web application (c) and obtains the corresponding graphics content from the database through interface (c.3). Depending on the request, the server can act in two ways:

- the low-resolution version of the graphical content is directly transmitted to the client (b.4), and/or
- additional modifications are performed on the server on the high-resolution version of the content: scene concatenation, 3D encoding, visualization, frame capturing and video encoding. Several output options are supported for the video stream: MPEG-4 ASP [12] or Flash [9], with low or high compression quality. The video stream is saved in the database and URL is transmitted in a SMS (trough Skype [24]) or as an e-mail (b.6) to the mobile recipient (b.5).

4 Results

Fig. 5. shows several snapshots of the mobile client illustrating the initial page, the animation recording page and the character selection page.

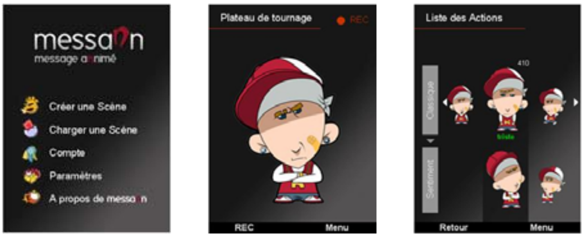


Fig. 5. Snapshots of the mobile application

Table 1. Transmitted data size in different scenarios

Number of animations	Using background	Using animation stored on the server	Data transfer [Kbyte]
1	**No**	**No**	**160**
1	**Yes**	**No**	**205**
1	**No**	**Yes**	**240**
1	**Yes**	**Yes**	**300**
3	**No**	**No**	**160**
3	**Yes**	**No**	**205**
3	**No**	**Yes**	**335**
3	**Yes**	**Yes**	**380**
10	**Yes**	**Yes**	**750**

One of the main barriers in using the mobile phone for advanced data communication application remains the cost per transmitted byte. By using only low-resolution models and synchronization between the client and the server through content IDs, we minimize the amount of data to be transmitted. Table 1 shows different usage scenarios and the price in terms of total bandwidth and total production time. The tests are done with list of 5 avatars, 5 animations per avatar and 5 backgrounds.

If LD files are used, the creation time is from 3-10 sec. If HD files are used, the creation time is from 5-10 min.

5 Conclusion and Future Work

In this paper we presented a rich communication approach for short message in mobile environment. By assisting the mobile phone with internet servers, it is possible to transform it in a powerful terminal for 3D graphics auto-production. For the future, this kind of application can be upgraded to be able to add sound in messages, possibility to capture background with camera and use captured voice.

References

1. 3dsmax, 3D Studio Max, Autodesk. Webpage, `http://www.autodesk.com/3dsmax`
2. Arsov, I., Preda, M., Prêteux, F.: MPEG-4 3D graphics for mobile phones. In: Proceedings First International Workshop on Mobile Multimedia Processing (WMMP 2008), Tampa, FL (2008)
3. Badler, N., Phillips, C., Webber, B.: Simulating Humans: Computer Graphics, Animation, And Control. Oxford University Press, Oxford (1993)
4. Bartels, R.H., Beatty, J.C., Barsky, B.A.: An Introduction To Splines For Use In Computer Graphics &Amp; Geometric Modeling. Morgan Kaufmann Publishers Inc., San Francisco (1987)
5. Blanz, V., Vetter, T.: A Morphable Model For The Synthesis Of 3D Faces. In: Computer Graphics Proceedings, Annual Conference Series, ACM SIGGRAPH, pp. 187–194 (1999)
6. Chadwick, J.E., Haumann, D.R., Parent, R.E.: Layered Construction For Deformable Animated Characters. Computer Graphics (SIGGRAPH 89 Conference Proceedings) 23(3), 243–252 (1989)
7. Decarlo, D., Metaxas, D., Stone, M.: An Anthropometric Face Model Using Variational Techniques. In: Proceedings of The 25th Annual Conference on Computer Graphics And Interactive Techniques SIGGRAPH 1998 (1998)
8. Dooley, M.: Anthropometric Modeling Programs –A Survey. IEEE Computer Graphics And Applications 2(9), 17–25 (1982)
9. FLV/F4V Technology Center, `http://www.adobe.com/devnet/flv`
10. Hilton, A., Beresford, D., Gentils, T., Smith, R., Sun, W.: Virtual People: Capturing Human Models To Populate Virtual Worlds. In: Proceedings of The Computer Animation, May 26-28, p. 174. IEEE Computer Society, Washington (1999)
11. Hirose, M., Deffaux, G., Nakagaki, Y.: Development Of An Effective Motion Capture System Based on Data Fusion And Minimal Use of Sensors. In: VRST 1996, ACM-SIGGRAPH And ACM-SIGCHI, pp. 117–123 (1996)
12. ISO/IEC JTC1/SC29/WG11: Standard 14496 2, a.k.a. MPEG 4 Part 2: Visual, ISO (1999)

13. Lander, J.: Over My Dead, Polygonal Body. Game Developer Magazine, 1–4 (May 1999)
14. Maestri, G.: Digital Character Animation 2: Essential Techniques, New Riders, Paperback (July 1999)
15. Maya, Autodesk, `http://www.autodesk.com/maya`
16. Menache, A.: Understanding Motion Capture For Computer Animation And Video Games, 1st edn. Morgan Kaufmann Publishers Inc., San Francisco (1999)
17. Molet, T., Boulic, R., Thalmann, D.: Human Motion Capture Driven By Orientation Measurements. Presence: Teleoper. Virtual Environ. 8(2), 187–203 (1999)
18. NASA Man-Systems Integration Standard (NASA-STD-3000), Revision B (July 1995)
19. NASA Reference Publication 1024, The Anthropometry Source Book, Volumes I And II (1978)
20. Preda, M., Prêteux, F.: The OLGA project: how standards, enable on-line gaming over, heterogeneous, networks. In: Proceedings 13th International Conference on Systems, Signals and Image Processing (IWSSIP 2006), Budapest, Hungary, September 2006, pp. 1–2 (2006)
21. Scheepers, C.F.: Anatomy-Based Surface Generation For Articulated Models Of Human Figures. PhD Thesis, The Ohio State University, Adviser: Richard E. Parent (1996)
22. Seo, H., Yahia-Cherif, L., Goto, T., Magnenat-Thalmann, N.: GENESIS: Generation Of E-Population Based On Statistical Information. In: Proceedings Of The Computer Animation, CA, June 19 - 21, p. 81. IEEE Computer Society, Washington (2002)
23. Singh, K.: Realistic Human Figure Synthesis And Animation For VR Applications. PhD Thesis, The Ohio State University, Adviser: Richard E. Parent (1995)
24. Skype, `http://www.skype.com`
25. Tamagotchi, `http://en.wikipedia.org/wiki/Tomagutchi`
26. Thalmann, N.M., Thalmann, D.: Complex Models For Animating Synthetic Actors. IEEE Computer Graphics And Applications 11(5), 32–34 (1991)
27. Vollee, `http://en.wikipedia.org/wiki/Vollee`
28. Waters, K.: Modeling 3D Facial Expressions: Tutorial Notes. State Of The Art In Facial Animation. ACM SIGGRAPH, 127–160 (1989) (SIGGRAPH 1989 Course Notes #22)
29. Watt, A., Watt, M.: Advanced Animation And Rendering Techniques. ACM, New York (1991)

On-Line Animation System for Learning and Practice Cued Speech

Ivica Arsov, Blagica Jovanova, Marius Preda, and Françoise Preteux

Télécom & Management SudParis (ex INT)
9 rue Charles Fourier, 91011 Évry Cedex, France
{Ivica.Arsov,Blagica.Jovanova,Marius.Preda}@it-sudparis.eu,
Francoise.Preteux@it-sudparis.eu

Abstract. This paper presents a set of technologies developed with the goal to improve the learning and practice of Cued Speech (CS). They are based on 3D graphics and are covering the entire end-to-end content chain: production, transmission and visualization. Starting from the requirements of an online system for CS, the research and development path took into account real-time constraints, personalization, user acceptability and equally important, the easiness and feasibility of the deployment. The original components of the system include 3D graphics and animation encoders, streaming servers and visualization engines and are validated in two applications: a web service for text to animation conversion and a chat service supporting two or more users.

Keywords: computer graphics, cued speech, avatar animation, real time transmission, 3D graphics player, MPEG-4 standard.

1 Introduction

The most effective and most used mean of communication between people is the spoken language. When the faculty of hearing is lost, people suffer of social isolation. Education of impaired hearing children is very difficult and depends on the availability of signers, physically accompanying the child in the classroom and translating the speech in Sign Language (SL). An alternative to the sign language is the Cued Speech [6] that consists mainly in lips reading together with hand motion that allows for better distinction between sounds with the same lips shape. While CS is less difficult to learn and practice than the SL, this visual mean of communication still requires the physical presence of a signer having face contact to the audience.

The huge steps achieved in technology development in the last decades, both in hardware and software, create today the premises of several computer mediated visual communication methods. The traditional approach is based on natural video (filming a real person) and an alternative is by using synthesized images from speech or written text. The importance of integration of people with disabilities in a society is signalized by the European legislation that requires that access to all services is made equally available to all citizens. This means that a very large number of companies, offering very diverse services, will have to address the issue of communication with sensory impaired people. At present, such communication relies heavily upon human sign language interpreters, but there can never be enough of these skilled individuals to be present at every face-to-face interaction or even to sign a large proportion of broadcast television. Nowadays only some television programs are "translated" by incrusting in the video frame a SL signer filmed in studio conditions. Teaching CS or SL in interactive applications is a big challenge some solutions being currently available. The traditional manner in teaching them is by using highly illustrated books[1] sometimes accompanied by filmed material. With the development of Internet and its transformation into a multimedia channel, several video web servers, especially the ones supplied by the community (Web2.0) include content for discovering or learning CS or SL. As an example Youtube contains several CS channel proposing more than 50 video lessons. Even more pedagogical is the on-line service developed by Michigan State University[2] which provides the video representation in American Sign Language for more than 4500 English words. By selecting a specific word, the video containing a person signing the word is presented. Unfortunately, it is not possible to compose complete sentences or to input any word. Producing the natural video data (for SL or CS) is costly and time consuming because it requires the presence of the signer and does not allow reuse and repurpose of the filmed content. Such limitations conducted to an increasing interest for synthetic images. By producing the images from text or speech, the synthetic video content become as easy to manipulate as the text itself [13]. A direct application consists in the use of film textual subtitles for enriching the movie with a synthesized signer. By associating a speech to text converter, some systems (e.g. TELEFACE, SYNFACE) allow phone conversation between deaf and hearing persons [4, 12]. In these systems, speech information from telephone is processed at the receiver side and an animated face is created. Similar concepts were developed in the TELMA project, supported by a software and hardware system with audiovisual functionalities for a telecommunication terminal (cellular phone) [12] and in the system described in [1]. More interactive scenarios including exercises for practicing CS are implemented in BALDI [5]. However, the signed content is pre-recorded being impossible to synthesize new words or sentences.

In this paper we are presenting an online system, able to synthesize in real time face and hand animation for CS, based on the text or speech inputted by the user. The novelty of our approach consists in the system architecture based on a dedicated server able to perform costly operations and to deliver the results as a compressed animation stream. On the user side, only a 3D graphics player is required, being possible to implement it on light terminals. Such approach has the advantage of processing the speech very close to the capture place, avoiding lost of the voice quality due to transmission errors and bandwidth. The proposed architecture is illustrated in Fig. 1. On the server side, the two entries, voice and text, are converted in animation parameters. The latter are encoded as an MPEG-4 animation stream and broadcasted to the network. On the client side, an MPEG-4 player receives the animation stream and updates, in a continuous manner, the scene graph defining the

[1] http://www.cuedspeech.org/sub/resources/learning.asp

[2] http://www.easycartsecure.com/LanguageMatters

avatar's face and hand. The design of the system is based on a set of technical and non-technical requirements. Table 1. summarizes the technical requirements and gives an overview on how we addressed them.

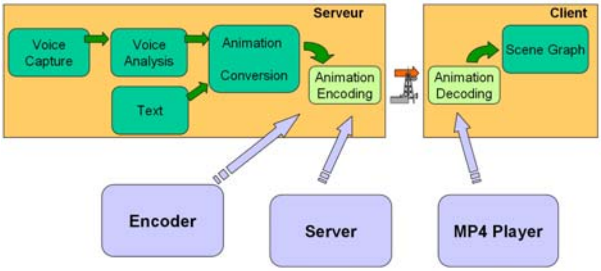


Fig. 1. In the proposed architecture, the transmission is performed for animation parameters.

Table 1. Requirements for on-line CS learning and practicing system.

Requirement	Proposed Solution
Synthesize visual content from text and/or speech, in real time	A synthetic video (graphics) based solution is preferred to a natural video one.
Voice analysis should be performed in optimal conditions	Since this module is very sensitive in terms of the quality of the input and requires significant processing power, in the proposed system it is implemented on the server side.
Transmission data should be very compact	The transmitted data consists in animation parameters. To make them compact MPEG-4 for avatar animation is used.
Support several clients in the same time	The transmission protocol is UDP
The software implemented on the client should be light enough to be able to run on mobile phones and PDA	By using the MPEG-4 standard, the only software running on the client is a multimedia MPEG-4 player (including 3D graphics capabilities)
The user may change the representation of the avatar	The avatar is locally loaded in the player by reading an MPEG-4 file describing the avatar geometry and appearance.
Same animation computed by the server should be usable by different avatars	A protocol for creating compatible avatars was established.

The first requirement conducts toward a computer graphics model. A video based solution, while presenting the advantages of signing quality and user adoption, does not allow easy manipulation, signs composition and repurposing. The rest of the paper is organized as follows. Section 2 surveys the face animation approaches. Section 3 introduces the avatar animation model of MPEG-4 and Section 4 describes in details the implemented system. We conclude the paper and present the perspectives in the last section.

2 Computer Graphics Face Animation

The process of facial animation is a very challenging task: on one hand the human face is composed of muscles that need to be perfectly coordinated in order to look realistic; on other hand, the human eye is very sensitive to facial movements, so it can notice even the slightest inconsistency and it may interpret the image in different way. While complex, the face and general human like animation attracted researchers from early times of computer graphics. Since extended surveys on face animation already exist (e.g. Deng and Neumann's book, Data-Driven 3D Facial Animation [8]), we only highlight the main approaches for visual speech animation and complete the survey with a detailed presentation of the MPEG-4 approach based on morphing space, the latter being less presented in the literature.

Starting from the pioneering work of Parke [19], an entire set of tools from computer graphics were proposed, adapted and tested for face animation. They may be classified in tools performing updates directly on the mesh or tools based on deformation controllers, which parameterize the mesh space, evolve in time and transfer their evolution to the mesh. The deformation controllers are trying to simulate face muscles with different forms such as lines [3], splines [17, 23, 24] or free form deformations [7, 14]. In more complex muscle approaches, it is possible to animate them based on physics properties [25]. One of the most known CS animation system is BALDI, based on the geometrical facial model described in [16].

In an on-line scenario, supposing streaming, a strong requirement is the compactness of the animation parameters. The approaches based on parameterization of the face space or of the deformation controller are thus more appropriate. One of the most common methods is derived from FACS (Facial Action Coding System) [10] that includes a fixed number of action units, stored as a code. The code is transmitted to the player, interpreted and the local face is animated accordingly. A similar implementation was standardized in the first version of MPEG-4, specifying 84 Feature Points, two high-level animation parameters for visemes and expressions and 66 low level parameters for 3D displacements of the feature points. The complete animation is then performed by deforming the mesh in the vicinity of the key points [9, 11, 15]. While easy to control, this model is highly dependent on the quality of the parameterization which is very difficult to transfer from one face to another. For this reasons, the MPEG community introduced in 2004 a more flexible approach, based on a morphing space, which is described in details in the next section.

In CS animation the major part of the communication signal concerns the shape of the mouth. In addition to the spatial properties of the mesh deformation, in the animation of the spoken language, the temporal behavior is of main importance. In a stationary approach, we may assume that each phoneme has associated a specific viseme. In a dynamic approach, the phoneme, thus implicitly its visual representation, the viseme, depends on its neighbors in the sentence. Thus it is necessary to consider not unique phonemes, but a family of them, used depending on the context, technique known as co-articulation. Several studies demonstrate the importance of the sound and its synchronization with the animation in CS applications [2], especially for subjects with partial hearing capabilities. This conducts to the need of producing time stamped animation data that, within an online scenario, ensures the synchronization on the receiver side. The system requirements described above conduct naturally to a solution similar to the audio/video on-line systems (supporting streaming and synchronization), and the large set of functionalities of the MPEG-4 standard ensure some of them. However, the Face Animation tool as published in 1998 [18] has serious limitations [21] with respect to the number of face configurations. To overcome these limitations, the authors of this paper proposed for standardization the morph space tool. This technique, published as international standard in 2004 [21] is described in the next section.

3 MPEG-4 Tools for Avatar Animation

The MPEG-4 standard, unlike the previous MPEG standards, does not only cope with highly efficient audio and video compression schemes, but also introduces the fundamental concept of media objects such as audio, visual, 2D/3D natural and synthetic objects to make up a multimedia scene. As established in July 1994, the MPEG-4 objectives are focused on supporting new ways (notably content-based) of communicating, accessing and manipulating digital audiovisual data [20]. Thus, temporal and/or spatial behavior can be associated with an object. The main functionalities proposed by the standard address the compression of each type of media objects, hybrid encoding of the natural and synthetic objects, universal content accessibility over various networks and interactivity for the end-user. In order to specify the spatial and temporal localization of an object in the scene, MPEG-4 defines a dedicated language called BIFS - Binary Format for Scenes. BIFS inherits from VRML the representation of the scene, described as a hierarchical graph. In terms of functionalities related to virtual characters, both VRML and MPEG-4 standards define a set of nodes in the scene graph to allow for a representation of an avatar. However, only the MPEG-4 SNHC specifications deal with streamed avatar animations. A major difference is that an MPEG-4 compliant avatar can coexist in a hybrid environment and its animation can be natively synchronized with other types of media objects, while the H-Anim avatar can only exist in a VRML world and must be animated by VRML generic, usually non-compressed, animation tools.

Now that the reasons of virtual character standardization within MPEG-4 become clearer, the question is how to find the good compromise between the need for freedom in content creation and the need for interoperability? What exactly should be standardized, fixed, invariant and in the mean time, ideally imposing no constraints on the designer creativity? The long term experience that MPEG community has makes it possible to formulate a straight and solid resolution: in the complex chain of content producing-transmitting-consuming the interoperability is ensured by only standardizing the data representation format at the decoder side. Pushing this concept to its extreme, an MPEG ideal tool is that one for which two requirements are satisfied: the designer can use any production tool he/she possesses to create the content and it can be possible to build a full conversion/mapping tool between this content and an MPEG compliant one. The same principle has been followed when MPEG released the specifications concerning the definition and the animation of the virtual characters, and specifically human avatars: there are no “limits” on the complexity of the avatar with respect to its geometry or appearance and no constraints on the motion capabilities.

The animation method of a synthetic object is strongly related to its definition. A simple approach, is to consider the virtual character as a hierarchical collection of rigid geometric objects called segments, and to obtain the animation by transforming these objects with respect to their direct parents. The second method, called BBA

(Bone-based Animation) in MPEG-4 consists in considering the geometry of the virtual character as a unique mesh and to animate it by continuously deforming its shape using a one-dimensional controller. While the former offers low animation complexity with the price of the seams at the joints between the segments, the latter ensures a higher realism of the representation, but requires more computation. A third approach is to model the animation space by using target shapes (i.e. eigen vectors) and obtaining the animation as a linear combination of them. Since in CS, only the face and a hand are used, we are introducing in the next paragraphs the MPEG-4 tools for modeling and animating them.

The straightforward manner of representing the geometry of a face model and in general of the entire avatar is by using an indexed set of planar surfaces (usually triangles) obtaining by grouping 3D vertices (Fig. 2.). By attaching properties like colors, texture coordinates and normals on top of each vertex, it is possible to represent any 3D object. For complex objects, the sub-sampling in planar surfaces is denser thus conducting to significant amount of information to transmit or store. In order to reduce this data, MPEG-4 standardized several compression algorithms exploiting the spatial redundancy [20]. To animate such geometric models, two of MPEG-4 technologies were used: the first simulates the biomechanical properties of a skeleton-based system, and the second is based on the animation space built on target shapes and interpolation between them.

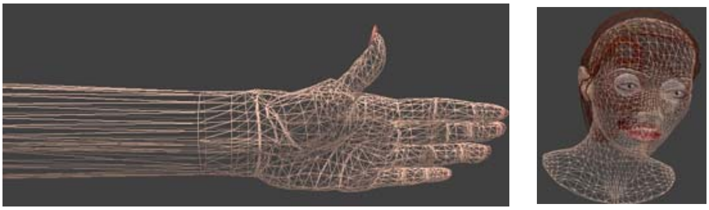


Fig. 2. Hand and Face of the CS virtual coder

Bone-based Animation. An arbitrary 3D mesh can be associated to a hierarchical structure of deformation controllers. The latter controls portions of the mesh by using a reduced number of parameters. In order to reproduce the biomechanical effects or the real articulated objects, the deformation controller may be as simple as a segment line (a bone). Grouping the bones into a hierarchical structure and transferring the geometric transformation of each bone to the vertices under its influence, produces the animation. A vertex on the mesh may be influenced by one or several bones. In the first case the geometric transformation of the bones is entirely applied to the vertex, in the second a weight factor between several influences is used. The new position of a vertex V_i influenced by several bones, each one transformed by T_k is obtained as follows:

$$V_i = \sum_{k=1}^{n_k[V_i]} \mu_k^i * T_k * (T_k^0)^{-1} * V_i \qquad (1)$$

with μ_k^i the weight factor between the bone k and the vertex V_i and T_k^0 is the initial transformation of the bone k (in the static position). The transformation T_k is expressed in global coordinates and is obtained for each bone as the product between its local transformation and the global transformation of its parent:

$$T_k = T_k^{local} \cdot T_{parent(k)} \tag{2}$$

Animating a model consists in updating the transformation T_k^{local} which for 3D space represents a matrix of 16 elements. Since for articulated organic objects only the rotation component of this matrix is generally used, MPEG-4 proposed decomposition in elementary components (translation, rotation and scale). Thus, the animation stream contains for each animation frame only the elementary transformations that are updated during the frame. This form conducts to very low compression bitrates. Let us note that the MPEG-4 standard does not impose any skeleton configuration neither on the initial pose, neither on the number of bones. By introducing T_k^0 in equation (1) the standard ensures consistency for arbitrary articulated object.

Morphing. The principle of morphing animation consists in representing each animation frame as a linear combination of target shapes (defining a basis[3] of an algebraic space). This technique is very low level, being independent on the deformation model. Thus the target shapes can be obtained by using arbitrary deformation tool applied on a neutral (static) version of the object. Once different configuration obtained, the basis can be computed by performing PCA (Principal Component Analysis). During the animation, V_i is computed as follows:

$$V_i = V_i^{static} + \sum_{k=1}^{n} w_k * (V_i^{\mathrm{target}_k} - V_i^{static}) \tag{3}$$

with $V_i^{static}$is the initial position of the vertex V_i, $V_i^{targetk}$ is the position of the same vertex in the target mesh k and w_k is the weight of the target mesh k in the current animation frame. Animation is obtained by simply updating the value of the w_k for all the meshes of the basis. Face animation by using morphing is a well known approach and the tool is supported by the majority of authoring tools. The MPEG-4 standard does not impose any limit with respect to the dimension of the morph space.

The two MPEG-4 animation approaches presented above were designed with the goal of minimizing the animation data. For skeleton based animation, the local transformation of bones (generally a rotation) is transmitted and for morphing based animation, a weight factor for each target shape is transmitted. To reduce even more the quantity of data and to support streaming, MPEG-4 standardized a specific syntax of the compressed animation frame. In a similar manner as the video encoding, animation encoding is using temporal prediction, signal de-correlation, quantification and entropy encoding. We refer to [22] for the description of the encoding algorithm and compression results.

[3] Not always the target shapes define a basis in the mathematical sense of the term, since the orthogonality is not always ensured.

4 Implementation and Simulation Results

Based on the technology presented in the previous section, we developed and integrated an end-to-end content chain for producing, transmitting and visualizing CS content. In the following we describe in details each component.

Production. Here, there are two components: off-line preparation of the avatar and on-line animation generation from speech and text. The first is based on a protocol establishing the way to build the configurations for the face and for the hand. The output is an MPEG-4 file defining the avatar. In our model, the face is defined by 11 target shapes (Fig. 3.), there are 9 configurations for the hand and 6 positions.

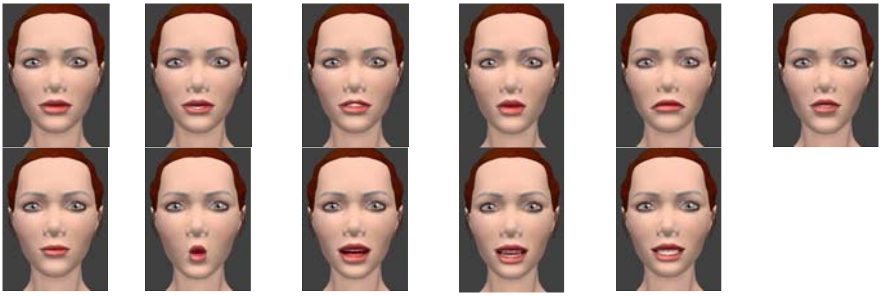


Fig. 3. Different target shapes defining the morph space

The online animation production exploits the previously obtained MPEG-4 file and for each phoneme, converts it into MPEG-4 animation parameters. These parameters are then compressed by the BBA encoder.

Transmission. The output of the BBA encoder provides animation data encapsulated in standard Access Units (AU) [26]. Each AU contains time information that can be used for obtaining transport packets. Two transport protocols are currently implemented: UDP and RTSP.

Visualization. On the client size, we developed an MPEG-4 3D Graphics player able to load a local or remote file or stream, decode the geometry, the texture and the animation, and rendering the 3D graphics scene.

Based on the components presented above we developed two prototypes for learning and practicing CS: a web service where the user input the text and visualize the animation, and a chat service, allowing a CS communication between two users. In both prototypes server-client architecture is used, where the server has the role of computing and encoding the MPEG-4 animation parameters, and the client has only the role of decoding and visualizing the avatar. Together with the animation, a sound track is computed (synthesized from text), compressed (in MPEG-4 AAC), transmitted and played by the same MPEG-4 player. Since similar approaches are used in both prototypes[4] we are only presenting in details the components of the Chat service. The architecture of the Chat service for two clients is illustrated in Fig. 4.

The two main components of the system are the Application Server (AS) and the Chat Client (CC). The AS is managing the communication between the clients and the conversion from text to speech and animation. The CC is getting the input from the user, sends it to the AS as ASCII text, decodes and displays the animation and the audio received from the server. The AS is composed of the following units: *Chat server*, *Text to speech and CS engine*, *CS to BBA convertor*, *WAV to AAC converter* and *RTSP streaming server*. The *Chat server* is responsible for managing the communication session (login, authentication) and the text exchanges between the clients. The text messages received from the clients are first sent to the *Text to speech and CS engine*, and when the synthesized audio and animation are ready for streaming, it sends the text to the other clients. The *Text to speech and LPC engine* is used to convert the text to synthesized audio and CS commands for animation.

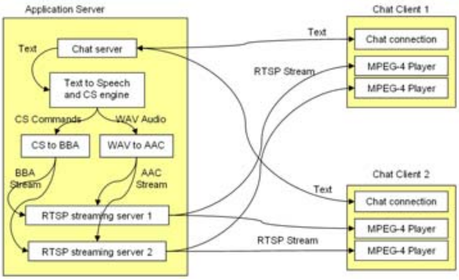
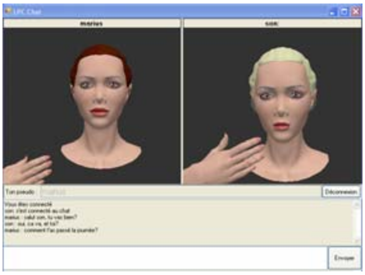


Fig. 4. The architecture of the Chat service



Vous êtes connecté
son: s'est connecté au chat
marius : salut son, tu vas bien?
son : oui, ca va, et toi?
marius : comment t'as passé la journée?

Fig. 5. Chat client interface

The CS data and the audio are converted into more usable streams. The CS is converted into a BBA stream, which can be used directly by the MPEG-4 player to display the animation. The audio is compressed into an AAC stream. Then both bitstreams are sent to the RTSP servers, which in turn send them to the appropriate Client. The use of the RTPS server is needed to achieve synchronization between the animation and the sound. There is a separate RTSP server for each connected user. Therefore the chat client initializes a new RTSP server for each new connected user. The Live555 library[5] is used to implement the RTSP server and it was extended to be able to read and synchronize the BBA and AAC streams. The CC is composed of the following units: *Chat connection* (Fig. 5.) and *MPEG-4 player*. The process of connecting to the Chat server is performed in the following manner: after the user authenticates, he joins the chat-room and he can send and receive text. Additionally the user receives from the server the parameters for initializing the RTSP connection. The client initializes the MPEG-4 player that loads a local MPEG-4 file containing the avatar. After connection to the streaming server, it is ready to receive animation

[4] The web service is available on-line at www.MyMultimediaWorld.com in the section "Learn LPC".

[5] http://www.live555.com/liveMedia/

and audio data. When the user enters text, the text is sent to the server, and all the connected users receive the entered text, the animation and the audio stream.

5 Conclusion

Starting from a set of technical requirements we designed and implemented several key components and an end-to-end architecture able to offer practical systems for learning and practicing Cued Speech in on-line environments. The approach, consisting in computing server and light (multimedia) clients was validated in two prototypes, a web and a chat service. The measurements have shown the ability of the system for supporting real-time interaction, low bandwidth communication and low-complexity on the user terminal. Our perspective is to extend the service for mobile clients, where stronger requirements with respect to the available bandwidth and terminal capacity are present.

Acknowledgment

We would like to thank French Research Agency ANR for financing LABIAO and INRIA/LORIA for speech recognition component, EDF R&D and DATHA for the work on establishing the requirements and the content production protocol.

References

1. Aboutabit, N., et al.: TELMA: Telephony for the Hearing-Impaired People.From models to User Tests (2007)
2. Albrecht, I., Haber, J., Seidel, H.-P.: Speech synchronization for physics-based facial animation. In: Proceedings of WSCG 2002, pp. 9–16 (2002)
3. Beier, T., Neely, S.: Feature-based image metamorphosis. In: SIGGRAPH proceedings, pp. 35–42. ACM Press, New York (1992)
4. Beskow, J., Dahlquist, M., Granström, B., Lundeberg, M., Spens, K.-E., Öhman, T.: The Teleface project - Multimodal Speech Communication for the Hearing Impaired. In: Eurospeech 1997, Greece (1997)
5. Caplier, A., Bonnaud, L., Malassiotis, S., Strintzis, M.G.: Comparison of 2D and 3D Analysis For Automated Cued Speech Gesture Recognition (2004)
6. Cornett, R.O.: Annotated bibliography of research on Cued Speech. CS Journal 4, 77–99 (1990)
7. Coquillart, S.: Extended free-form deformation: A sculpturing tool for 3d geometricmodeling. Computer Graphics 24, 187–193 (1990)
8. Deng, Z., Neumann, U.: Data-Driven 3D Facial Animation. In: Softcover, illus. in color, vol. VIII, 296, p. 149 (2008) ISBN: 978-1-84628-906-4
9. Doenges, P., Lavagetto, F., Ostermann, J., Pandzic, I.S., Petajan, E.: MPEG-4: Audio/video and synthetic graphics/audio for mixed media. Image Communications Journal 5(4) (1997)
10. Ekman, P., Friesen, W.V.: Facial Action Coding System. Consulting Psychologists Press (1978)

11. Escher, M., Pandzic, I.S., Thalmann, N.M.: Facial deformations for mpeg-4. In: Proc. of Computer Animation 1998, Philadelphia, USA, pp. 138–145. IEEE Computer Society, Los Alamitos (1998)
12. Faulkner, A.: SYNFACE: A speech-driven synthetic face as a communication aid for hearing-impaired people. ELSNews 10.3, 3 (2001)
13. Gibert, G., Baily, G., Elisei, F.: Evaluation d'un système de synthèse 3D de Langue françaice parlée complétée, in Les Journées d'Étude sur la Parole, JEP (2006)
14. Kalra, P., Mangili, A., Thalmann, N.M., Thalmann, D.: Simulation of facial muscle actions based on rational free from deformations. In: Eurographics, vol. 11, pp. 59–69 (1992)
15. Lavagetto, F., Pockaj, R.: The facial animation engine: Toward a high-level interface for the design of mpeg-4 compliant animated faces. IEEE TCSVT 9(2), 277–289 (1999)
16. Massaro, D.W.: Symbiotic Value of an Embodied Agent in Language Learning. University of California, Santa Cruz (2004)
17. Nahas, M., Hutric, H., Rioux, M., Domey, J.: Facial image synthesis using skin texture recording. Visual Computer 6(6), 337–343 (1990)
18. Ostermann, J.: Animation of synthetic faces in mpeg-4. Proc. of IEEE CA (1998)
19. Parke, F.I.: Computer generated animation of faces (1972)
20. Pereira, F., Touradj, E.: The MPEG-4 Book. IMSC Press, Upper Saddle River (2002)
21. Preda, M., Prêteux, F.: Critic review on MPEG-4 Face and Body Animation. In: Proceedings IEEE International Conference on Image Processing (ICIP 2002), Rochester, NY, vol. 3, pp. 505–508 (2002)
22. Preda, M., Jovanova, B., Arsov, I., Prêteux, F.: Optimized MPEG-4 animation encoder for motion capture data. In: 3D Web Technology (Web3D 2007), Perugia, Italy, pp. 181–190 (2007)
23. Viad, M.L., Yahia, H.: Facial animation with wrinkles. In: Proceedings of the Third Eurographics Workshop on Animation and Simulation (1992)
24. Wang, C.L.Y., Langwidere, D.R.F.: A new facial animation system. In: Proceedings of Computer Animation, pp. 59–68 (1994)
25. Waters, K.: A muscle model for animating three-dimensional facial expression. In: SIGGRAPH Proceedings, vol. 21, pp. 17–24 (1987)
26. ISO/IEC JTC1/SC29/WG11: Standard 14496 2, a.k.a. MPEG 4 Part 2: Visual (1999)

AUM and Enterprise Tomography: New Concepts for Technology Enhanced Learning for Enterprise Systems in Higher Education

Liane Haak, Jan Aalmink, and Dirk Peters

Carl von Ossietzky University Oldenburg,
Department of Computer Science, Business Information Systems I / VLBA
Ammerländer Heerstrasse 114-118, 26129 Oldenburg, Germany
{Liane.Haak,Dirk.Peters}@uni-oldenburg.de
Ent.tomo@googlemail.com

Abstract. Modern teaching methods in the field of applied computer science can not ignore the teaching of well know application and information systems, like enterprise systems, e.g. ERP-Systems. Therefore case studies are the most chosen way to introduce stepwise the handling of these systems. Effective teaching concepts have to improve this situation by consideration of pedagogical and didactical aspects which supports the individual learning process of each student. Our actual research considers actual needs of higher education e.g. present learning in a lab as well as e-learning courses supported by new methods in technology enhanced learning by recording student's behaviour to guide him through the system. Therefore we introduce a concept using AUM and Enterprise Tomography to improve the teaching for Enterprise Systems.

Keywords: Technology Enhanced Learning, Enterprise Systems, Higher Education, Application Usage Mining, Web Usage Mining, E-Learning, Enterprise Tomography.

1 Introduction

Considering different perspectives in computer science and business economics technology aspects are highly complex in many areas and results to many problems. That one reason why Enterprise Systems (ES) it selves are getting more and more important for educational environments. But this interference and complexity are making the teaching and learning in this field quite difficult. For this reason practically approved systems are needed to give the students the chance to get their own practical experience with these technologies. Enterprise Systems have the capability for future pedagogic innovation within higher education. Their potential results from the possibilities in illustration, visualization and simulation of business and decision-making processes to students [1]. The main goal of using Enterprise Systems as e.g. Enterprise Resource Planning (ERP) Systems, Business Intelligence and others in higher education is to prepare the students for real work life and to give them practical experience in the application of these technologies. Another objective is focused by software developing companies like SAP® or Microsoft®, e.g the students should get in touch with their special products as early as possible, so that they already know these products if they need to work with it later or are in the position to decide about investments. Our interests in research belong to the variety of ERP technologies and how they could be used to teach our students in the university environment. Hereby the focus lays besides the actual standard software on new concepts resulting from up to date research, e.g. Federated ERP (FERP) Systems [2].

This contribution will show how new learning concepts supported by technologies like Application Usage Mining (AUM) [3] and Enterprise Tomography (ET) [4] could improve the teaching of Enterprise Systems in higher education.

2 Background

Confronted with the global market and knowledge transfer which increasingly could have seen as the key driver of wealth creation, universities have to reinvent themselves. European higher education institutions are collaborating in developing competitive curricula to compete in an in increasingly commercial and international market. They face a major new challenge as knowledge providers in individuals' lifelong learning which means for universities to offer flexible easy access: they need more open technological solutions. Information and communication technologies (ICTs) offer the prospect of flexible learning services. User needs in terms of range, access and costs could be supported while adopting innovative learning processes. New technologies allow effective and cost-efficient learning tailored to the needs of individuals and it is a critical success factor for universities in their role as "knowledge centres" for lifelong learning. That is a reason why universities of tomorrow must be more flexible, different from those of today. Within the beginning virtualisation learners want to access personalised and flexible learning services, available as and when they want them, as part of a continuum of lifelong learning.

For teaching Enterprise Systems like ERP systems the usage of case studies is a very common method. Therefore leading software companies like SAP® or Microsoft® provide a customized environments in form of "test" companies (e.g. the International Demonstration and Education System (IDES)[1] by SAP® and the Cronus AG[2] by Microsoft® accompanied by teaching material in form of case studies. Additionally SAP® build up an own higher education competence network with the University Competence Center (UCC) in Magdeburg and Munich based on the SAP University Alliance®. Nevertheless mostly the teaching it based on material used for software training courses. This allows a good overview and sometimes detailed impress of function within the systems. But one problem is the missing background about company structures and processes and other concurrent interests which should be considered and the missing of pedagogic and didactic aspects. For using case studies it is important that the learners develop their own solutions in a strategic way and make reasonable decision [5]. The case study design is already a critical success factor, as you can see in figure 1:

[1] http://help.sap.com/saphelp_46c/helpdata/DE/af/fc4f35dfe82578e10000009b38f839/frameset.htm

[2] http://www.microsoft.com/germany/bildung/infopool/mbsaa.mspx

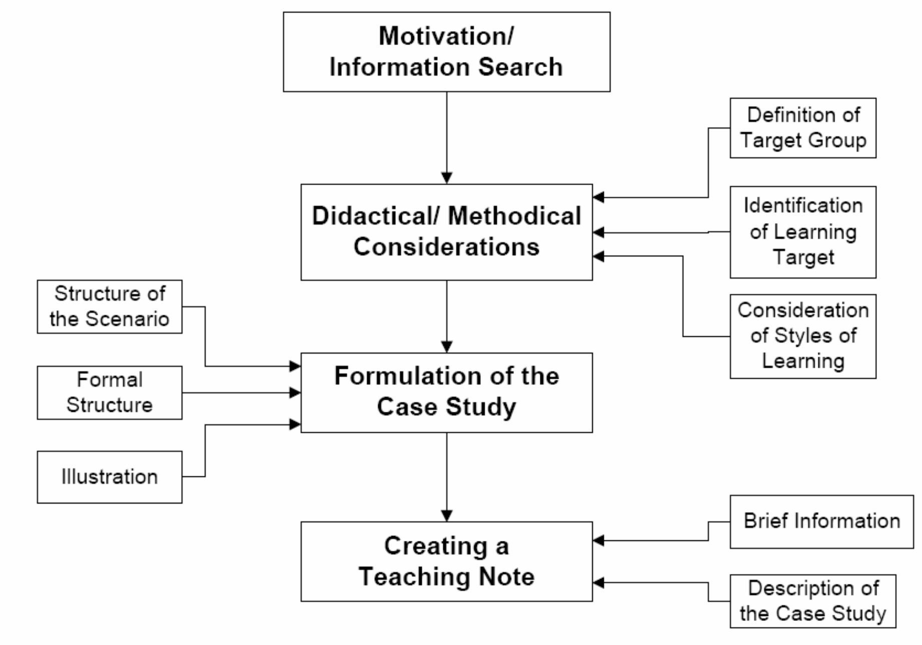

Motivation/ Information Search
Didactical/ Methodical Considerations
Formulation of the Case Study
Creating a Teaching Note
Definition of Target Group
Identification of Learning Target
Consideration of Styles of Learning
Structure of the Scenario
Formal Structure
Illustration
Brief Information
Description of the Case Study

Fig. 1. Conception of case study design [5]

The authors describe a concept for case study design in the environment of higher education using an example from the data warehouse area [5]. And it is obvious that case studies offer the possibility for an activity-oriented education with the focus on self-directed learning. But we still doubt if the correct case study design could already improve the situation. So there are still problems occurring from this missing background information and the difference of domain knowledge of the students. The material doesn't reflect the different education and major studies of the students; it is the same for all, anyway if they are e.g. economics or computer scientist. Our idea is to use technologies in the wide area of Technology Enhanced Learning (TEL) to improve these education methods and to guide the learner through his learning process.

3 Technology Enhanced Learning

Often Technology Enhanced Learning and E-Learning are used in the same way, but there are some differences. E.g. the term of Technology Enhanced Learning is the wider expression, which means that TEL is focusing on the technological support of any pedagogical approach that utilizes technology. That is maybe one reason why the existing definitions for Technology Enhanced Learning broadly spread and change continuously due to the dynamic nature of this evolving research field. Anyway, the definition of TEL must be as general as possible in order to capture all aspects:

TEL encompasses virtual and physical technology enhanced learning environments (incorporating physical learning spaces, institutional virtual learning environments, personalized learning environments and mobile and immersive learning environments). And the aim is to explore and develop effective practice in the delivery of flexible, seamless and personalised services to learners, focussing on the technological interface between the learner and their learning environment[3]. Learning activity consists of learning resources, actions, context, roles and the learning objective to support the learner to his learning goals, respecting individual as well as organizational learning preferences. Technologies play an important role in supporting these activities. That is surely one reason why the European Union (EU) support a number of project in this area, e.g. within in the 6th framework the Network of Excellences PROLEARN[4] and Kaleidoscope[5] which have the main objective to have shape the research area around TEL. In this research environment we want to introduce our new concept for teaching Enterprise Systems in higher education.

4 Concept for TEL for Enterprise Systems in HE

For our research we choose an approach in the field of Application Usage Mining (AUM) by Gamal Kassem which could be used to analyze the user's behaviour in Very Large Business Applications like SAP ERP for example [3]. Related to the Web Usage Mining (WUM), the AUM uses the transaction data, which is recorded during the usage process of an Enterprise System in adequate log-files, to generate specific behaviour patterns. The results of the AUM can be applied for the optimization of business processes, etc. In our approach, we want to use this algorithm to identify the knowledge level, behaviour patterns or possible deficits of a user in handling an Enterprise System. These results can be used to develop user specific exercises, depending on his/her skills. In this way it is possible to have an adaptive case study, which is built on the starting position of one individual student and which doesn't need to cover a heterogeneous group of students.

4.1 Application Usage Mining

When we are talking about the characteristics of today's business application systems, the focus is mainly on the integration of different enterprise sectors in order to have an all-round product, which covers all the functionality a modern enterprise is dealing with. These business application systems – especially ERP Systems – are dealing with the integration of data, functions and processes of all components of an enterprise. As a negative side-effect of this development, the user faces a rising complexity and an enormous amount of opportunities to navigate through these systems. Thus it is getting more and more difficult for users to handle such complex systems in an appropriate way. One approach which tries to solve this problem is the already mentioned Application Usage Mining (AUM) approach introduced by Kassem in [3]. This AUM approach uses the traces a user leaves in the system, via log-files and

[3] http://www.jisc.ac.uk/whatwedo/programmes/elearning/tele.aspx

[4] http://www.prolearn-project.org/

[5] http://www.noe-kaleidoscope.org/pub/

protocols for example. The idea is based on the methods of the widely spread Web Usage Mining, which has been successfully executed in the field of web applications. Kassem identified several significant differences between web applications and business applications systems, which legitimate this new approach. First of all, the web application access is anonymously, while a user in a business application system must be identified. Furthermore a visitor of a web site does not need an authorization, whereas the user of a business application system has a certain authorization, according to his function in the enterprise in order to perform certain tasks. A third difference which becomes apparent while comparing AUM with WUM is the fact, that the user's behaviour and the objective is completely free or undefined, while a user of a business application system has to perform predefined tasks according to a business process, which have to be executed efficient and with an optimal performance [6]. The complete list of the attributes, which can be used to describe the differences between web applications and business applications for WUM and AUM is shown in table 1.

Table 1. Comparison of WUM and AUM [6]

Attribute	**Web Application**	**Business Application System**
System access	Anonymous	The user must be identified
Authorization	No authorization necessary	The user has a certain authorization according to his function in the enterprise to perform certain tasks
Protocol	Standard http	No standards
Software	Web application – usually based on html-documents	Different software platforms are available
User's behaviour	Free	Execution of predefined tasks and business processes
User's objective	Undefined	Optimal performance of tasks and business processes
Application's aim	There is a possibility to reach a lot of visitors and customers	Efficient execution and automation of business processes

Other approaches like Data Mining, Process Mining or Workflow Mining are also playing a role in this context. Data Mining can be seen as a basis of all the approaches mentioned above. Process Mining for example is a specific type of Data Mining: Data Mining focuses on the creation of knowledge out of large data amounts and Process Mining uses therefore the knowledge of processes. Process Mining extracts data, which describe the execution of processes in order to model, save and reuse the process knowledge [7]. Workflow Mining is used to control, improve, execute and monitor business processes [8]. Therefore Workflow Mining aims on extracting information about processes from transaction logs for the purpose of visualizing the current status of a workflow model.

4.2 AUM for TEL

How can we use the AUM approach in the TEL environments? As described above, the AUM approach can be used to identify specific behaviour patterns of a user or maybe possible deficits a user shows during the work with an Enterprise System or a business application. We have a vision of a learning system, which for example can be used in the curricula of universities in order to teach students in handling a certain enterprise system like SAP ERP for example. We want to face the problem of the existence of heterogeneous learning groups, different knowledge levels and divergent learning behaviours by setting up a system, which dynamically generates user-specific tasks on the basis of the user patterns classification. A possible example could be the situation that a user needs too much time to complete a certain task or step in a business process, which is supported by the enterprise system; it might be possible to give him a hint in the form of a possible solution for his problem. The solution is not only based on a technical analysis of his behaviour via log-files or transaction data, it should also be possible to integrate didactical approaches into the solution finding algorithm in order to find a suitable way for the user to complete his exercise or task. Moreover it is favoured, that the given solution will be analyzed in a further step for the purpose of optimizing the solution-giving process. The result of this analysis should be integrated in the following hints or solutions the learning system gives to the user. In this way we can assure a continuous improvement of our system.

In the following section we will point out, how the concept of Enterprise Tomography could improve this concept in the field of Very Large Business Applications. The usage of the Enterprise Tomograph, which is introduced in the next chapter, can also be part of the idea of improving the TEL environments in the field of higher education.

5 Enterprise Tomography Supporting Adaptive Learning Environment

In general, real world VLBAs and Enterprise Business Process Platforms are heterogeneous software conglomerates with a high degree of business logic integration. Comprehension of business logic requires understanding of integration concepts in Enterprise Information Systems. Typically, integration knowledge is taught based on educational materials in combination with training and demo-systems. Integration concepts, coded in an Enterprise Information System and their utilization are normally not tangible and therefore cannot be measured. This implies that most training systems analyze the learning progress of an individual trainee based on Q&A sheets at the end of a training course. This, of course is detached from his real learning behaviour. With the integrated Enterprise Tomography approach, the phase of determining and evaluating the extent of learning progress can be streamlined and automated. Furthermore, the training scenario, as a deployable artefact, can be adapted in the course of the training life-cycle.

To be more precise, the learning metric is calculated and evaluated on a continuous basis and therefore the learning paths of training map can be optimized depending on learning metrics, which are indicators for the individual learning progress. E.g. if an individual trainee gets stuck in a business process path, he will receive immediately more detailed instructions to circumvent the impasse. Contextual training material in combination with learning metrics provides a big potential of efficiency along the training life-cycle in the learning map. The phase Q&A becomes more and more obsolete with a holistic approach based on Adaptive Learning Environments. The calculation of learning metrics is based on actual operative data taken into consideration by business process usage mining.

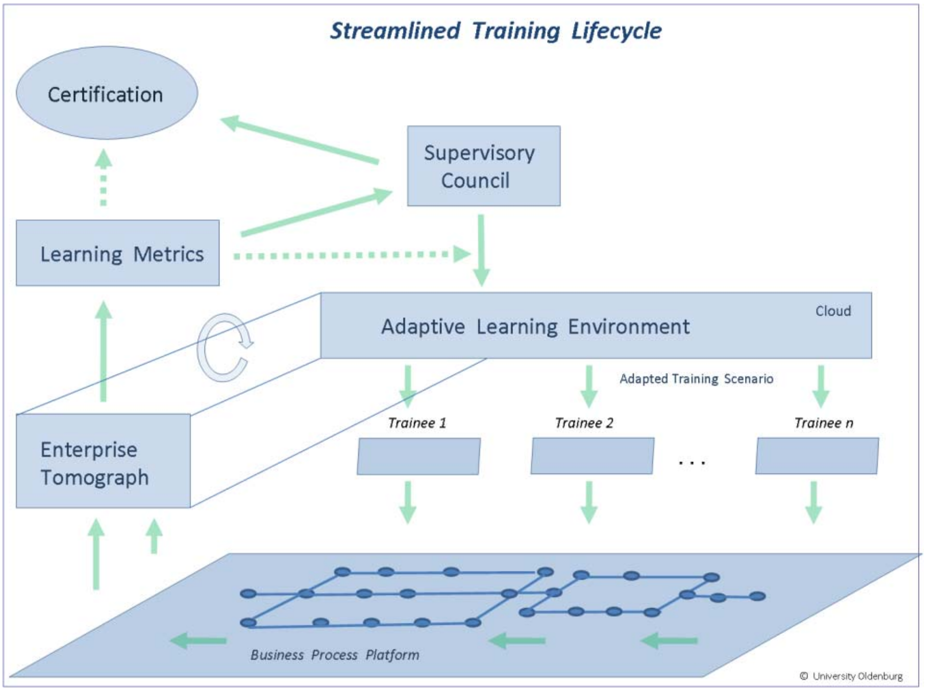


Fig. 2. Streamlined Training Life-Cycle with continuously adapted training scenarios

Figure 2 illustrates the holistic approach with real-time learning-progress analysis. It is essential, that the training life-cycle complies to the closed-loop paradigm. The Business Process Platform is continuously crawled by the Enterprise Tomograph. The Enterprise Tomograph can be seen as a domain-specific semantic search engine for full and delta-spaces. As shown in figure 2, a business process is a composition of business process steps orchestrated in parallel threads and sequences. Each individual process step has one or more Business Objects assigned. A business process step transforms the set of assigned Business Objects from one consistent state to the next according to the transactional ACID-Principle. In this context we have a semantic bijection between Business Objects as meta-data, object instances, ontologies and integration concept instances. Polymorphic integration concepts are the 'DNA' of Very Large Business Applications and Business Process Platforms regardless of the architectural and technical representation [9].

The Enterprise Tomograph takes the integration concepts into consideration and provides time and space-efficient generic operations on the set of integration concepts. Especially, the Delta Operator of the Enterprise Tomograph can calculate the difference of two sets of integration concept instances. To put it more specific, the Delta Operator determines the difference between system state A and B at point of time t0 and t1 projected on a semantic view. The delta in this context represents the progress of an individual trainee executing a specific business process of his individual learning map. The delta is the footprint on the database that is a good basis for derivation of learning metrics. Having such delta as a unit in place, we can easily provide the inverse operator, i.e. undo functionality. The trainee gets with help of the inverse operator his business process rolled back and can restart the business process from the beginning. Counting of restarts is one aspect of the learning metrics.

Each business process step can be regarded as a semantic milestone along the learning path. The Delta Operator calculates the last position in the business process graph. The last position can be regarded as a semantic milestone and is another aspect of learning metrics. Based on learning metrics, the Adaptive Learning Environment can decide dynamically and adapt the current active training scenario. E.g. if a trainee gets stuck at a specific business process step, he will get more detailed instructions for performing the business process step. The semantic context to be provided can be derived from the delta provided by the Enterprise Tomograph. If the learning metrics complies to quality standards, a work-flow can be triggered for the finalization of the certification procedure. Attached to the metric based certificate, a tomogram of the Enterprise Tomograph might be a good basis documenting the learning curve. In our scenario we have shown and reused the generic concept of Enterprise Tomography in the context of Adaptive Learning Environments. Basically, the Enterprise Tomography approach is an efficient interdisciplinary Application Usage Mining approach for Enterprise Platforms and Very Large Business Applications (VLBA). Enterprise Tomography semi-automatically identifies and localizes semantic integration concepts and visualizes integration ontologies in semantic genres.

Especially delta determination of integration concepts is performed in dimension space and time. Enterprise Tomography in general supports software and data comprehension. Large scaled Adaptive Learning Environments can benefit from the new paradigm based on mathematical algorithms. Adaptive Learning Solutions needs to support the training life-cycle. The learning behaviour and the learning progress are to be evaluated with scientific algorithms. One task of an adaptive learning solution is to calculate learning metrics of trainees or trainee groups using a VLBA or an heterogeneous composed Business Process Platform as an on-premise training system. The Adaptive Learning Environment is predestined for cloud computing, because it is a non-mission critical application.

In accordance to apparatus-supported medical diagnostics, an Enterprise Tomograph scans a Business Process Platform and its business data. Utilizing concept mining algorithms, integration ontology trees are extracted from the training system and its contextual data [9], [15]. These forests are indexed with standard algorithms originating from bio informatics [10], [12], [13]. An in memory based domain-specific search engine, containing in the Enterprise Tomograph, provides refinement techniques and enables a time-efficient localization of integration ontologies. The center of attention is a generic delta operator being implemented by the Enterprise Tomograph [4]. The delta operator determines the delta of integration ontologies between point of time tn and tn+1 (dimension time) or the delta of integration ontologies between two system instances (dimension space). The delta can be projected on an area of expertise categorized in semantic categories. Standard tree mapping algorithms visualize the delta [11], [14]. The Enterprise Tomograph takes the holistic data universe into consideration ranging from the VLBA software, business data, and metadata to contextual data. In this extended data universe integration concepts are evaluated and visualized.

6 Conclusions and Future Work

At the beginning of this paper we described the necessity of effective teaching concepts in the field of Enterprise Systems in higher education. We pointed out, that teaching methods have to considerate pedagogical and didactical aspects to support every individual student or learner in his learning process. Based on the new methods in the field of Technology Enhanced Learning we discussed the Application Usage Mining approach as one possible technical fundament to create a suitable learning system or surrounding. Taking the behaviour patterns, the AUM generates from a user's behaviour in a business application, it could be possible to offer the learner a dynamic user-specific environment to get a better idea how today's complex business application systems work. Furthermore we described the Enterprise Tomograph, which also can be used to analyze the marks a user leaves in a system. The existing prototype can already be used to make conclusions about different steps a user passes in a business process.

The future work will be about the improvement of the theoretical concepts. We will develop architectures, which will show the interacting elements in a more detailed way. After creating this theoretical framework it is possible to develop a prototype for AUM in TEL, where the Enterprise Tomograph can act as a basis or an input-giving element in order to awake the Application Usage Mining idea in Technology Enhanced Learning environments.

References

[1] Ask, U., Juell-Skielse, G., Magnusson, J., Olsen Dag, H., Päivärinta, T.: Enterprise Systems as Vehicles of Pedagogic Innovation - Enterprise System Inclusion in Higher Education. In: Proccedings of the 5th International Conference on Enterprise Systems, Accounting and Logistics (5th ICESAL 2008), Crete Island, Greece, July 7-8 (2008)

[2] Brehm, N., Haak, L., Peters, D.: Using FERP Systems to introduce web service-based ERP Systems in Higher Education. In: Abramowicz, W., Flejter, D. (eds.) Business Information Systems Workshops (BIS 2009), Poznan / Poland (April 2009)

[3] Kassem, G.: Application Usage Mining: Grundlagen und Verfahren. Shaker Verlag (2007)

[4] Aalmink, J., Marx Gómez, J.: Enterprise Tomography - an efficient approach for semiautomatic localization of integration concepts in VLBAs. In: Cruz-Cunha, M.M. (ed.) Social, Managerial and Organizational Dimensions of Enterprise Information Systems (2009) ISBN: 978-1-60566-856-7

[5] Hans, D., Marx Gómez, J., Peters, D., Solsbach, A.: Case study-design for Higher Education – a demonstration in the data warehouse environment. In: Abramowicz, W., Flejter, D. (eds.) Business Information Systems Workshops (BIS 2009), Poznan / Poland (April 2009)

[6] Kassem, G., Marx Gómez, J., Rautenstrauch, C.: Analysis of User's Behaviour in Very Large Business Application Systems with Methods of the Web Usage Mining - A Case Study on SAP® R/3®. In: Menasalvas, E., Segovia, J., Szczepaniak, P.S. (eds.) AWIC 2003. LNCS (LNAI), vol. 2663, pp. 329–338. Springer, Heidelberg (2003)

[7] Schimm, G., van der Aalst, W., ter Hofstede, A., Weske, M.: Mining most specific Workflow Models from Event-based Data. In: van der Aalst, W.M.P., ter Hofstede, A.H.M., Weske, M. (eds.) BPM 2003. LNCS, vol. 2678, pp. 25–40. Springer, Heidelberg (2003)

[8] Zur Mühlen, M., Hansmann, H.: Workflowmanagement. In: Becker, J., et al. (eds.), pp. 374–407 (2005)

[9] Abels, S., Haak, L., Hahn, A.: Identification of common methods used for ontology integration tasks. Interoperability Of Heterogeneous Information Systems. In: Proceedings of the first international workshop on Interoperability of heterogeneous information systems, Bremen, Germany, pp. 75–78. ACM, New York (2005)

[10] Abouelhoda, M.I., Kurtz, S., Ohlebusch, E.: Replacing suffix trees with enhanced suffix arrays. Journal of Discrete Algorithms 2, 53–86 (2005)

[11] Bille, P.: A Survey on Tree Edit Distance and Related Problems. Theoretical Computer Science 337(1-3), 217–239 (2005)

[12] Dementiev, R.: Algorithm Engineering for Large Data Sets, Doctoral Thesis, University of Saarland, Saarbrücken, pp. 151–156 (2006)

[13] Kärkkäinen, J., Sanders, P., Burkhardt, S.: Linear Work Suffix Array Construction. Journal of the ACM (J. ACM) 53(6), 918–936 (2006)

[14] Lu, C.L., Su, Z.-Y., Tang, C.-Y.: A new measure of edit distance between labeled trees. In: Wang, J. (ed.) COCOON 2001. LNCS, vol. 2108, pp. 338–348. Springer, Heidelberg (2001)

[15] Nicklas, D.: Ein umfassendes Umgebungsmodell als Integrationsstrategie für ortsbezogene Daten und Dienste, doctoral thesis, University Stuttgart, Online Publication, Stuttgart, pp. 44–53 (2005)

Multiplatform Real-Time Rendering of MPEG-4 3D Scenes with Microsoft XNA

Sasko Celakovski[1] and Danco Davcev[2]

[1] ITGMA, Skopje, Macedonia
sasko.celakovski@itgma.com
[2] FEIT, Skopje, Macedonia
etfdav@feit.edu.mk

Abstract. This study presents approach for presentation of MPEG-4 3D graphics objects in real-time using Microsoft XNA technology. The proposed approach considers the aspects of real-time 3D rendering on multiplatform environment. We introduce management of MPEG-4 3D resources to address the rendering requirements of a modern real-time 3D rendering engine. In our approach this management results in extension by enabling appropriate representation of data resources. This study presents an example of using MPEG-4 encoded 3D content in advanced 3D visualization applications such as games and virtual reality on Windows and Xbox systems.

Keywords: MPEG-4, 3D Scenes, Rendering, XNA.

1 Introduction

Among the existent or on-going multimedia standards, MPEG-4 [1] is one of the most complete in terms of media representation, compression, 2D and 3D graphics primitives, user interaction and programmatic environment. As a member of the MPEG family, the MPEG-4 standard inherits and improves all the features of its predecessors, offering the possibility of efficient transmitting and/or storing a huge amount of digital audio / video. Furthermore, the standard addresses state of the art techniques such as advanced audio coding, video compression-based on visual object, wavelet deployment and mesh-based representation. In addition to representing elementary media, MPEG-4 goes further and specifies mechanisms that allow creating complex multimedia scenes. It is now possible to combine several media, to define synthetic content and to add interaction and dynamic behavior of the scene.

The features of MPEG-4 are supported by using a special description language called BInary Format for Scenes (BIFS). BIFS is based widely on Virtual Reality Modeling Language [2] (VRML) and represents a binary encoded version of an extended subset of VRML, which can represent roughly the same scene as with VRML in a much more efficient manner. In this paper we are interested in the BIFS functionalities for representing synthetic content, especially the 3D objects. The wide range of functionalities supported by MPEG-4 makes this standard one of the most complete and advanced solution, and companies are slowly deploying MPEG-4 technologies inside their applications. The complexity of the standard is a serious drawback and its wide acceptance as a common multimedia format has difficulties to take off.

Real-time rendering engines are used to visualize 3D content together with animations while instantly reacting to different user interactions. These engines are widely used in applications for entertainment, educations, GUI and so on. Today's modern 3D engines bring highly advanced design architectures capable of presenting complex 3D content in efficient and hardware independent manner. The technologies that enable high level of hardware abstraction layer (HAL) allow functioning of the rendering engines on a wide range of graphics hardware. Two most important HAL technologies are DirectX [3] and OpenGL [4].

In this study we presents an example of using MPEG-4 encoded 3D content in advanced 3D visualization applications. The proposed approach uses Microsoft's XNA [5] framework library as a HAL layer to enable multiplatform rendering of the MPEG-4 3D Scenes. The visualization architecture is described in detail in Section 2. The MPEG-4 scene hierarchy and access to the 3D content is explained in Section 3. In Section 4 we show how content decoding is integrated in the rendering engine. Section 5 concludes the paper.

2 Real-Time Visualization

Microsoft XNA framework library is a set of tools with a managed runtime environment provided by Microsoft that facilitates computer game development and management. XNA attempts to free game developers from writing repetitive code and bring different aspects of game production into a single system.

The XNA Framework is based on the native implementation of .NET Compact Framework 2.0 for Xbox 360 development and .NET Framework 2.0 on Windows. It includes an extensive set of class libraries, specific to game development, to promote maximum code reuse across target platforms. The framework runs on a version of the Common Language Runtime that is optimized for gaming to provide a managed execution environment. The runtime is available for Windows XP, Windows Vista, and Xbox 360. Since XNA games are written for the runtime, they can run on any platform that supports the XNA Framework with minimal or no modification. The XNA Framework thus encapsulates low-level technological details involved in coding a game, making sure that the framework itself takes care of the difference between platforms when games are ported from one compatible platform to another, and thereby allowing game developers to focus more on the content and gaming experience.

XNA framework library provides set of classes that managed different aspects of the complete functionality that a rendering engine should provide. The object oriented approach provides suitable code encapsulation of the functionality. In that way abstract representation of the 3D scene is provided that enables handling of different input formats.

XNA's Graphics namespace provides sufficient functional for presentation of 3D graphical entities. It defines the Model class as collection of 3D Meshes which store the visual attributes of the 3D objects and collection of Bones which are used for animation of the 3D objects. In this study we will only consider static 3D objects.

Each 3D object in the proposed rendering architecture is stored in a *Mesh* class. The *Mesh* class contains multiple *ModelMesh* objects that store the geometric and visual attributes of a complex 3D object. *ModelMesh* consists of the vertex buffer that stores the complete set of vertices that describe the object, index buffer that defines the complete set of triangles that describe every object in the mesh, bounding sphere that represents spherical approximation of the volume of the complete mesh and collection of effects providing the appearance attributes for the 3D object. Figure 1 presents the connections between these elements.

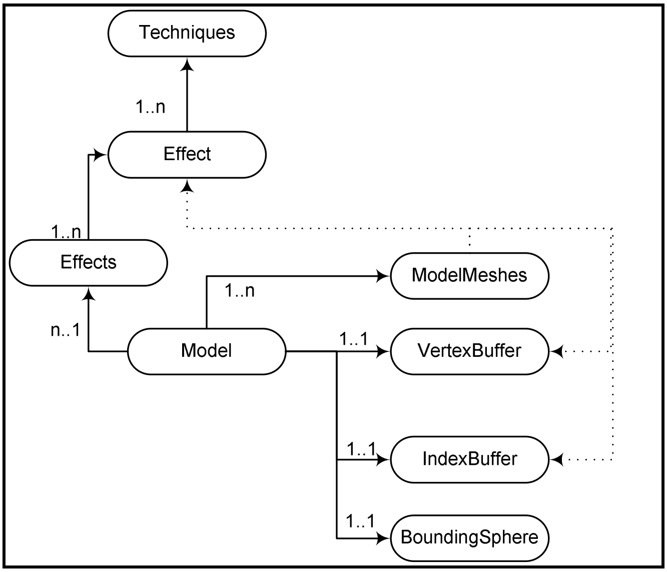


Fig. 1. Hierarchy of Classes for 3D Object presentation

Real-time visualization application would utilize these data structures to build optimal procedures and methods for visualization of 3D content. The rendering engine, as the set of these processes and methods are usually called, consists of hierarchical program classes that manipulate the data and hardware system function. Example diagram of such rendering engine is given in Figure 2.

The root object provides the basic initiation functionalities and communication with the operating system. Each instance of the application allocates only one Rendering System and one Scene Manager object. The Rendering System provides the functionality for management and utilization of the visualization hardware. Using the Microsoft's XNA framework and libraries, the Rendering System implementation in our experiment is multiplatform and can be handles graphical hardware available in standard PC configurations as well as Xbox console.

The Scene Manager class implements the functionality for manipulation of the models in the 3D scene. It handles the rendering data as well as global scene parameters.

It sorts the 3D objects from the scene for most optimal rendering performance, it selects the 3D objects that are inside the viewing (camera) space of the scene. Entity objects represents all of the entities in the scene, these can be 3D objects that are visualized.

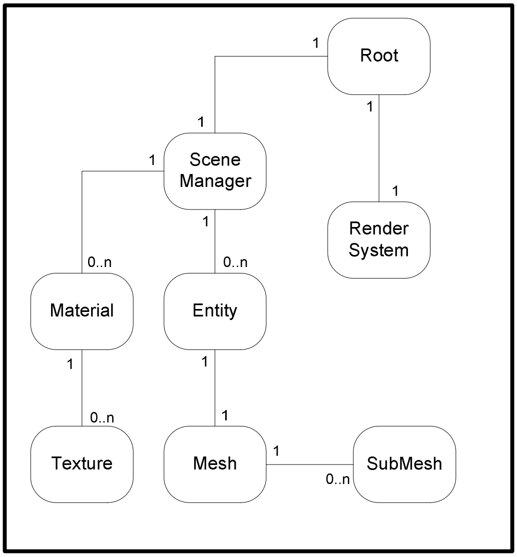


Fig. 2. Rendering Engine Diagram

3 MPEG-4 3D Scene

Creation of the 3D content (geometry, textures, animations etc.) is usually done in a 3D modeling tools (such as 3D Studio Max or Maya). The rendering engines commonly provide utilities capable to export 3D content into custom their formats. Our goal in this study is to propose a new source of 3D content based on the MPEG-4 3D standard, the .mp4 format, which can be effectively embedded into Real-time rendering engine based on XNA framework. This will allow application developers to facilitate advantages of the MPEG-4 3D compression and stream-ability into their applications. Content creation and generation of the .mp4 files is out of the scope of this paper and we assume that the 3D meshes are already created and stored in a proper MPEG-4 3D format.

In order to be able to visualize 3D meshed from the .mp4 file we need to be able to reconstruct their fundamental elements (vertices, triangles, materials etc.) from the compressed MPEG-4 format. For this purpose we need to have clear understanding of the internal scene structure of MPEG-4 3D scene.

The process of decoding of the MPEG-4 content file into data structures that are more suitable for visualization in real-time is presented in Figure 3.

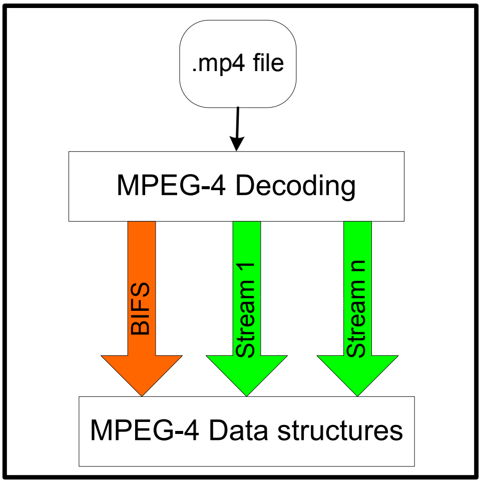


Fig. 3. Decoding MPEG-4 file

In order to feed the rendering engine through the data structures described in the previous section, we need to effectively decode and access the MPEG-4 BIFS. The elementary streams of BIFS contain the 3D nodes that should be visualized by the rendering engine. One implementation of this decoding is based on GPAC [6] open source MPEG-4 encoding/decoding library. GPAC is an open source project for advanced content storage and manipulation. It is a framework consisting of multimedia functionally written for research and academic purposes that covers different aspects of multimedia with focus on visual technologies (graphics, animation and interactivity). GPAC library implementation is compliant to the MPEG-4 standard and presents a small and lightweight alternative to the official MPEG-4 reference software implementation. It also provides robust 2D and 3D visualizing applications for MPEG-4 content. However, distinctive difference between the GAPC implemented viewers and our approach presented in this paper is our focus on multiplatform support and application in high performance graphical applications like games and real-time virtual environments.

In this experimentation we choose to implement the “Animated Character” profile [7]. This profile selects 11 nodes out of more than 150 MPEG-4 specified nodes and 2 elementary streams that are sufficient for representation of skeletal animation and static objects. The list of scene nodes and elementary streams from this profile are given in the Figure 4.

The Animated Character Profile provides sufficient set of information for presentation of complex static 3D objects and hierarchically animated 3D models. It is designed with purpose to present animated human 3D characters with complete set of geometrical, visualization and animation attributes. For the purposes of this paper, we have experimented with static 3D objects containing geometrical description of several thousands of polygons, material properties and texture map. Vertex data is stored in the Coordinates structure, the indexing of vertices into triangles is provided through IndexFaceset elements. Normal vectors are stored in the Normal structure

Animated Character Profile (MPEG-4)

List of Specific Nodes:

SBSkinnedMesh
Shape
IndexedFaceset
Coordinate
Normal
Appearance
Material
ImageTexture
MorphShape
SBVCAnimation
SBBone

List of Streams:

BBA
JPEG

Fig. 4. MPEG-4 Nodes and Streams in the Animated Character Profile

from the Animated Character profile. Visualization attributes are preserved in the Appearance, Material and ImageTexture data structures. Images used as textures, as referenced through the ImageTexture are stored in the JPEG stream. The experimentation developed for the purposes of this paper we have not considered the data available and stored in the following data structures from the Animated Character profile: MorphShape, SBVCAnimation and SBBone. Also, the data provided in the BBA stream, which contains the animation parameters for the bones of the animated character, is not being processed by the experimental programmed developed for the purposes of this paper.

4 Integration of MPEG-4 into XNA

The real-time visualization of MPEG-4 3D scene begins with decompression and parsing of the 3D objects stored in a compressed MPEG-4 file. After successfully decompressing the nodes from the Animated Character Profile, the information is stored ina data structure that should be adequately mapped to the data structures of the rendering engine.

The 3D static objects in the MPEG-4 scene are represented through instances of the *Shape* class and the 3D animated objects through instances of the *SBSkinnedModel* class. In this study we will consider only 3D static objects.

The *Shape* class contains information for the 3D geometry and its appearance. The geometry is given in MP4 *IndexedFaceset* structure that contains an array of vertices and indexed faces. The *Appearance* class gives the description of the appearance of the 3D geometry object. This information is used to generate the initial 3D scene.

Upon decoding of the MPEG-4 content the create scene process is initiated. This process populates the data structures for the rendering engine by creating the relevant object instances. Information from the *IndexedFaceset* is mapped to the *VersteBuffersx* and *IndexBuffers* classes of the rendering engine, while the information from the *Appearance* fills-in the *Effects* and *Techniques* structures. The actual mapping of MPEG-4 structures and rendering engine's structures is given on Figure 5.

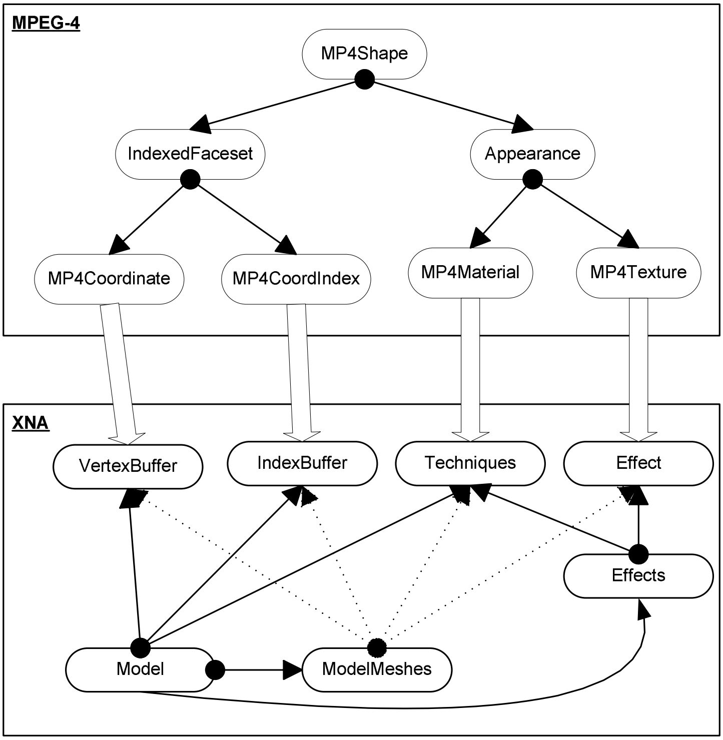


Fig. 5. Mapping of the 3D Scene from MPEG-4 to XNA data structures

Mapping of the data from compressed MPEG-4 source to the data structures suitable for fast access, browsing and real-time rendering is performed only once during the initiation of the application. Once that data is stored in the XNA structures and objects it is reused and manipulated from these structures and by methods provided within the Microsoft XNA framework.

5 Conclusion

In this study we have presented an approach for incorporating MPEG-4 defined 3D content in a commercial scale real-time rendering engine. Based on the MPEG-4 standard, the media entity (here the 3D objects) is first decoded and the MPEG-4 scene graph structure is filed. Then, using widely accepted framework for multiplatform rendering in real-time, the Microsoft's XNA, the decompressed 3D scene is visualized.

The proposed approach presents effective using of compressed 3D storage in various applications such as virtual worlds, games, educational and applications running in multiplatform environments on Windows and Xbox systems.

Our future work will be based on the integration of the elementary media (video, audio and animation stream) implying synchronization and stream control.

References

[1] ISO/IEC 14496-1:2001 Information technology – Coding of audio-visual objects – Part 1: Systems, International

[2] ISO/IEC 14772-1: Information technology — Computer graphics and image processing — The Virtual Reality Modeling Language — Part 1: Functional specification and UTF-8 encoding (1998)

[3] Microsoft DirectX, `http://www.microsoft.com/DirectX`

[4] OpenGL, `http://www.opengl.org`

[5] Microsoft XNA Framework, `http://msdn.microsoft.com/en-us/library/bb203940.aspx`

[6] GPAC Project on Advanced Content, `http://gpac.sourceforge.net/`

[7] Preda, M., Prêteux, F.: Virtual Characters in MPEG-4. IEEE Transaction on Circuits and Systems for Video Technology 14(7), 975–988 (2004)

Fast Classification Scheme for HARDI Data Simplification

V. Prčkovska[1], A. Vilanova[1], C. Poupon[2],
B.M. ter Haar Romeny[1], and M. Descoteaux[2]

[1] Dept. of Biomedical Engineering, Eindhoven Univ. of Technology, The Netherlands
{V.Prckovska,A.Vilanova,B.M.terHaarRomeny}@tue.nl
[2] NeuroSpin, CEA Saclay, France
maxime.descoteaux@gmail.com

Abstract. High angular resolution diffusion imaging (HARDI) is able to capture the water diffusion pattern in areas of complex intravoxel fiber configurations. However, compared to diffusion tensor imaging (DTI), HARDI adds extra complexity (e.g., high post-processing time and memory costs, nonintuitive visualization). Separating the data into Gaussian and non-Gaussian areas can allow to use complex HARDI models just when it is necessary. We study HARDI anisotropy measures as classification criteria applied to different HARDI models. The chosen measures are fast to calculate and provide interactive data classification. We show that increasing b-value and number of diffusion measurements above clinically accepted settings does not significantly improve the classification power of the measures. Moreover, denoising enables better quality classifications even with low b-values and low sampling schemes. We study the measures quantitatively on an ex-vivo crossing phantom, and qualitatively on real data under different acquisition schemes.

Keywords: High Angular Resolution Diffusion Imaging, Diffusion Tensor Imaging, Diffusion Weighted Magnetic Resonance Imaging, voxel classification, HARDI, DTI, DW-MRI.

1 Introduction

Diffusion tensor imaging (DTI) is a recent technique that can map the orientation architecture of neural tissues in a completely non-invasive way by measuring the directional specificity (anisotropy) of local water diffusion [1]. The diffusion tensor model however, has well known limitations in areas of complex intravoxel heterogeneity with crossing fibers, where the diffusion process can not be modeled as Gaussian. Nonetheless, DTI is still very popular and has many advantages like: fast and clinically feasible acquisition schemes (typically 7-60 number of gradients (NG), b-values 1000 s/mm^2 and total acquisition time of 3-5 minutes), fast post-processing of the data that allows interactivity in the data exploration, simple visualization techniques and modeling using well-developed tensor mathematics. To overcome the limitations of DTI, more sophisticated models were introduced using high angular resolution diffusion (HARDI). For HARDI, about sixty to a several hundred diffusion gradients are acquired in order to reconstruct certain spherical probability functions (SPFs) that either recover the underlying fiber populations or depict certain diffusion properties. Popular HARDI reconstruction techniques include ADC modeling [2,3], QBall imaging [4], diffusion orientation transform (DOT) [5], spherical deconvolution (SD) [6], and several other model-based methods. The produced output by the above techniques is always given in the form of a spherical function $\psi(\theta,\phi)$ that characterizes the local intra-voxel fiber structure. This function can be represented using spherical harmonics (SH)

$$\psi(\theta,\phi) = \sum_{l=0}^{l_{max}} \sum_{m=-l}^{l} a_l^m Y_l^m(\theta,\phi) \ , \tag{1}$$

where Y_l^m represent the spherical harmonics of order l and phase m, and l_{max} is the truncation order of the SH series.

HARDI has obvious advantages over DTI in crossing configurations, but has several drawbacks that come along with this complex modeling: longer processing time of the data (that can typically take few hours to a few days), inability to interactively explore the data because of over-cluttered and computationally heavy visualization as well as longer data acquisitions. Hence, one wonders if a complex high-order modeling of the data is always needed (i.e. at every voxel) or worths its drawbacks? In crossing areas, it is certainly justified, but for a large part of the white matter, there are significant single fiber voxels where high-order modeling might be redundant. Being able to classify regions of single fiber (Gaussian) and crossing fibers (non-Gaussian) in white matter in a fast and reliable way, is thus important. It can reduce the modeling complexity in areas where it is not needed enabling possibilities for data simplification that has many advantages for further post-processing and visualization od the data, especially w.r.t. reducing the computer memory requirements. This will undoubtedly make HARDI data easier to manipulate and interact with, making it more attractive for clinical applications.

There is a wide range of anisotropy measures proposed in literature, such as [2,3,4,7,8,9,10,11]. Several authors [3,8,9] have attempted to use some of these methods to classify non-Gaussian profiles, but all these attempts have been made on the apparent diffusion coefficient (ADC) profiles and without convincing real data results. In this paper, we explore and compare the classification power of these measures, and apply them on several different SPFs represented in SH basis. We also extract the number of maxima from the corresponding glyph representations. To illustrate the possible application of our data simplification from the classification output, we evaluate the gain in speed for calculation of 8^{th} order constrained spherical deconvolution (CSD) [12] only in non-Gaussian areas and 2^{nd} order ODFs calculated from diffusion tensors in Gaussian areas, compared to only use of 8^{th}-order CSD everywhere in the real data. An important contribution is that we study the classification measures under different b-values and sampling scheme acquisitions from several real HARDI datasets. We also validate the classification experiments on an *ex vivo* phantom with known ground truth. We thus come to several conclusions suggesting that HARDI processing and data interaction are possible in clinical settings.

2 Methods

We implemented several anisotropy measures from the literature, generalized anisotropy (GA) [10], generalized fractional anisotropy (GFA) [4], the Shannon entropy (SE) of [11], the cumulative residual entropy (CRE) of [7,8], as well as fractional multifiber index (FMI) [2], and R_0, R_2, R_i [9]. These measures were applied on the ADC profiles [2,3], analytical QBalls [13] and the DOT [5]. The DOT generally produces much sharper glyph profiles for high R_0 value at the cost of more noisy profiles with spurious peaks. Finding the *best* R_0 in real data is difficult and often done by observation [5]. Hence, to avoid this R_0 selection problem and inspired by definitions of the ODF from q-ball imaging [4] and the marginal ODF (mODF) from diffusion spectrum imaging (DSI) [14], we propose the similar ODFs computed from the DOT as:

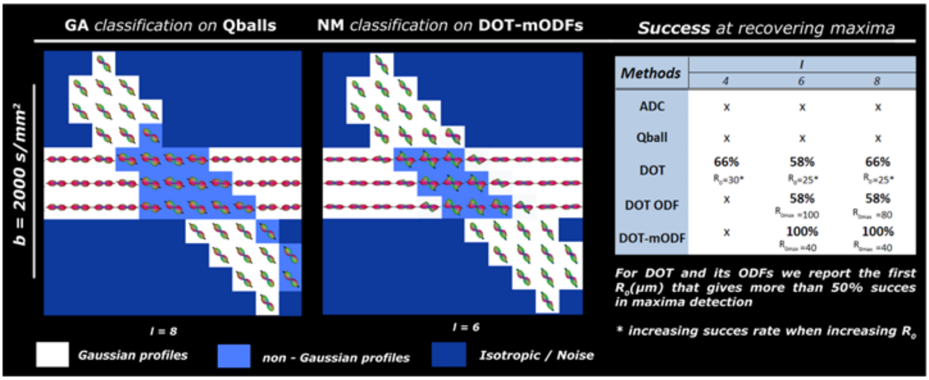
$$\psi_{\text{DOT-ODF}}(\theta,\phi) = \int_0^{R_{0max}} P(r,\theta,\phi)dr, \quad \psi_{\text{DOT-mODF}} = \int_0^{R_{0max}} P(r,\theta,\phi)r^2 dr, \tag{2}$$

where $P(r,\theta,\phi)$ is the probability density function (PDF) computed from DOT [5], and R_{0max} is set to a conservatively high value (as an example see table on Figure 1).

As a discrete binary measure for classification we propose to use the number of maxima (NM). NM uses the number of local maxima of the min-max normalized SPFs profiles, where the discrete spherical function surpasses a certain threshold (here, we use 0.6) from points on a fine discrete mesh. Moreover, for better visual perception, in our figures we generate min-max normalized RGB color coded glyphs, although one must keep in mind that this normalization enhances angular contrast of glyphs in the white matter but also deforms isotropic glyphs considerably.

Diffusion Data Acquisition ***Ex-vivo phantom:*** To test our classification measures, we use two real physical *ex-vivo* phantoms with fibre bundles crossing at 45° and 90° [15]. These datasets serve as ground truth, where the number of crossing and linear voxels is known. The phantom data was acquired on a 1.5T Signa MR system (GE Healthcare), TE/TR =130ms/4.5s,12.0s (45° and 90° phantom, respectively), BW=200KHz. We analyze the data acquired at two b-values of $b = 2000$ and $b = 8000$ s/mm^2, along 200 uniform directions.

Human: Diffusion acquisitions were performed using a twice focused spin-echo echo-planar imaging sequence on a Siemens Allegra 3T scanner, with FOV 208 × 208 mm, isotropic voxels of 2mm. Uniform gradient direction schemes with 49 and 121 directions were used and the diffusion-weighted volumes were interleaved with b_0 volumes every 12th scanned gradient direction. Datasets were acquired at b-values of 1000, 1500, 2000, 3000, 4000 s/mm^2 and in the same session,



Methods	l		
	4	6	8
ADC	x	x	x
Qball	x	x	x
DOT	66% R_0=30*	58% R_0=25*	66% R_0=25*
DOT ODF	x	58% R_{0max} =100	58% R_{0max} =80
DOT-mODF	x	100% R_{0max} =40	100% R_{0max} =40

For DOT and its ODFs we report the first R_0(μm) that gives more than 50% succes in maxima detection

** increasing succes rate when increasing R_0*

Fig. 1. Classification results from the phantom data

two anatomical data sets (192 slices, isotropic 1mm voxels) were acquired using the ADNI protocol for registration. Finally, before HARDI reconstruction of the ADC, QBalls and DOT, we applied a denoising pre-processing step [16], available online[1], to correct for the Rician noise bias in the datasets.

3 Results

Phantom Results: The 45 ° is a challenging angle where most of the HARDI techniques struggle to detect multiple maxima, especially at low b-values. We will first analyze the results from the maxima detection. As pointed in the work of Prckovska et al. [17], DOT has the potential of recovering small angles regardless of the b-value, which we demonstrate in the table of Figure 1. In the table we report the success at recovering two maximas in the crossing voxels by all of the examined SPFs. We additionally report the first R_0 for the DOT and its derivations at which the success is greater than 50%. Even more interesting, we observe that the derivation of the DOT discussed in Section 2, with its ODFs (DOT-ODF and DOT-mODF) manifest similar behavior as the DOT itself, which show a better angular resolution than QBall and suggest a better choice of reconstruction algorithm for fiber tracking purpose. The results from the NM classification on the 90 ° phantom are omitted, due to the 100% success demonstrated in all reconstruction methods.

For the rest of the anisotropy measures we can quantitatively describe the classification power of the 45 ° and 90 ° phantoms by using binary classification statistical test. We thus report the specificity and sensitivity of the classified crossing and linear voxels respectively. The sensitivity measures the proportion of actual positives which are correctly identified as such, and the specificity measures the proportion of negatives which are correctly identified. All the measures must be thresholded to obtain the classification and this process is sensitive. Two thresholds are needed to separate the interval of anisotropy values into three distinct compartments: Isotropic/noise, Gaussian and non-Gaussian. We

[1] http://www.irisa.fr/visages/benchmarks/

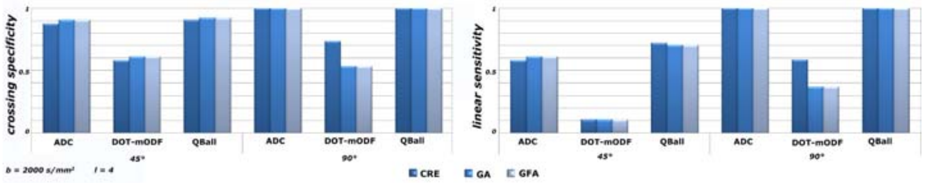


Fig. 2. Specificity and sensitivity demonstrated for the crossing and linear areas in the phantoms respectively.

thus iterate over the whole range of values of each anisotropy measure and find the interval where all the crossings are detected while the number of false positives stays minimal. For the purpose of data simplification, it is very dangerous the presence of false negatives (i.e. crossings detected as linear), because relevant information can be lost. Therefore to ensure absence of false negatives we set the sensitivity of crossing classification criteria to 1. In other words, we ensure that all the crossings voxels are always detected ($crossing\ sensitivity = 1$) and no crossing voxel is classified as linear ($linear\ specificity = 1$). In Figure 2, we present the specificity of the crossing classification for each measure and the sensitivity of the linear detected voxels for the 45° and 90° phantoms respectively. Any measure with high specificity is a good candidate for classifying the crossing regions.

We observe that all the three measures CRE, GA and GFA demonstrate similar classification power applied on ADC and QBall profiles. The DOT-mODF however (and similar DOT and DOT-ODF), is substantially worse, even though it produces sharper angular profiles. The sharper and thus, more spiky DOT profiles, are actually a disadvantage for classification measures, as they then produce many false positives in the linear voxels part of the branches of the phantom. Another important result is that increasing the order of SH representation does not significantly improve the classification power of the measures. The results coincide for the sensitivity of the linear classification. The other measures FMI, R_0, R_2 and R_i, demonstrate more irregularities and dependencies on the angular configuration of the crossing diffusion pattern and it was more difficult to find thresholds for classification. They were thus omitted in the tables but it is worth mentioning that they did improve at higher b-value when tested in the phantom dataset at $b = 8000s/mm^2$. Shannon entropy however, was impossible to threshold with our criteria and was omitted from the analysis. Therefore, from our *ex-vivo* phantom study, we can conclude that CRE, GA and GFA can be applied as a reliable classification between Gaussian and non-Gaussian profiles with in general less than 8% false positive classification results in any configuration. GA and GFA have advantage over CRE since they can be calculated only on the SH coefficients and therefore are significantly faster. Moreover, an SH order of 4 is sufficient to classify the non-Gaussian profiles. However, if one is interested in the number of maxima, it is then useful to use higher SH order to discriminate low angle crossings, such as 45°.

Human Data Results: The centrum semiovale was used to illustrate the qualitative analysis of the classification results. It is an interesting region for analysis,

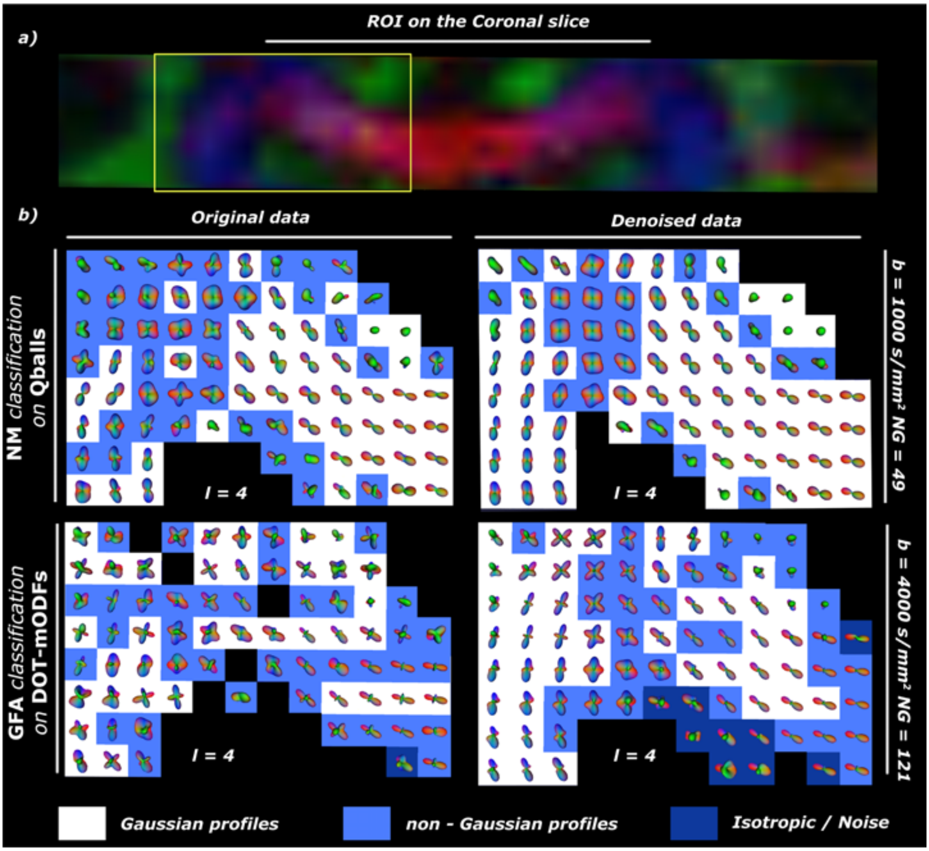


Fig. 3. The effect of denoising demonstrated on original versus denoised data in different acquisition schemes

since fibers of the corpus callosum (CC), corticospinal tract (CST), and superior longitudinal fasciculus (SLF) form different two-fiber and three-fiber crossing configurations in that area. The region-of-interest (ROI) was defined on a coronal slice (Figure 3a). It is important to mention that all the real data results are from similar regions, since those are different HARDI scans from the same subject, and have not been registered. We applied the same classification measures as for the phantom study on the original and denoised data from our datasets. Denoising dramatically improves the glyph profiles and the coherence of the non-Gaussian regions, as seen in Figure 3. We also observe a decrease in the irregularities in the crossing profiles. Our results suggest that even at low b-value, low NG and low estimation SH order, there is success in recovering crossing diffusion patterns and identifying linear regions. In opposite, going to very high b-values (i.e. $\geq 3000 s/mm^2$) and modeling the data with high SH order ($\geq$6) results in polluted glyphs regardless with or without a denoising

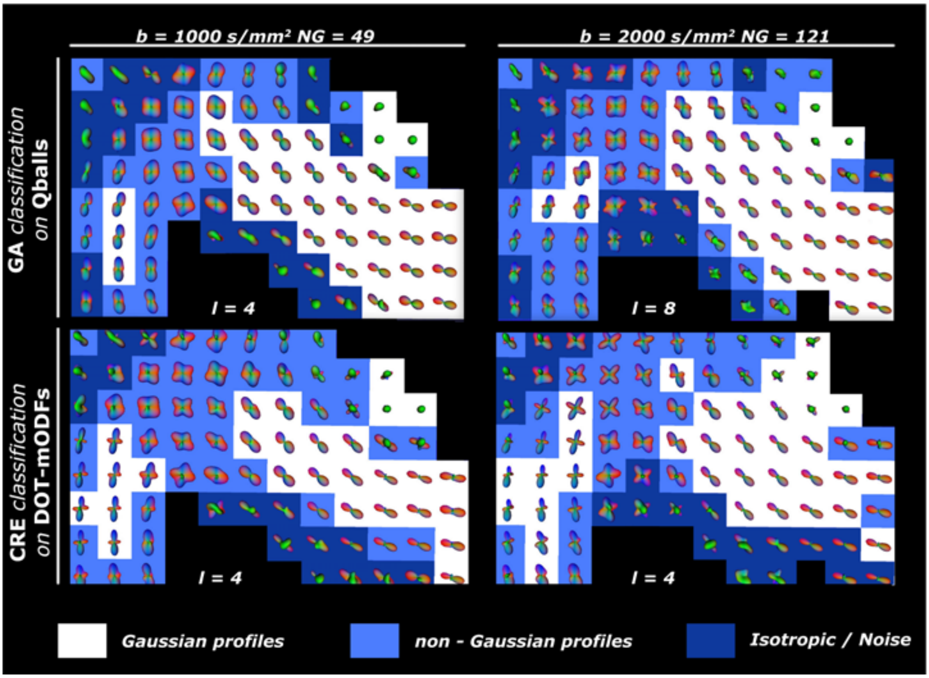


Fig. 4. Some examples from different classification

phase. Comparing the results of the classification from different measures, we observe that increasing the b-value sharpens the HARDI profiles and benefits only for maxima extraction purposes. However, there is no significant gain in classification of non-Gaussian profiles, as observed in the phantom study. This is seen in Figure 4, where we see sharper glyphs for DOT-mODF but similar classification power regardless the measure or acquisition scheme. We also note that increasing the model order ($l > 4$) does not increase the classification power. This leads to the conclusion that 49 directions is sufficient for recovering most of the crossings and non-Gaussian voxel detection, which can significantly reduce acquisition time (compared to a 121 NG acquisition).

As an example of possible application of our classification, in Figure 5 we show hybrid visualization of the simplified data (labeling provided by GA classification) from an *in-vivo* dataset represented with 8^{th} order CSD [12] in the non-Gaussian classified regions, and DTI ODFs in the Gaussian regions. The difference in running time is as follows: computing CSD of order 8 for the whole brain in white matter mask: 540 minutes (36601 voxels). Computing CSD of order 8 in labeled crossing : 120 minutes(8164 voxels). Computing DTI ODFs in labeled linear : 19 seconds[2]. With hybrid data modeling there is nearly a factor 5

[2] These times were calculated on a 1.66 GHz processor dual core Intel machine with 2 GB of RAM. The time can be improved by parallelizing the code and changing the parameters of CSD regularization.

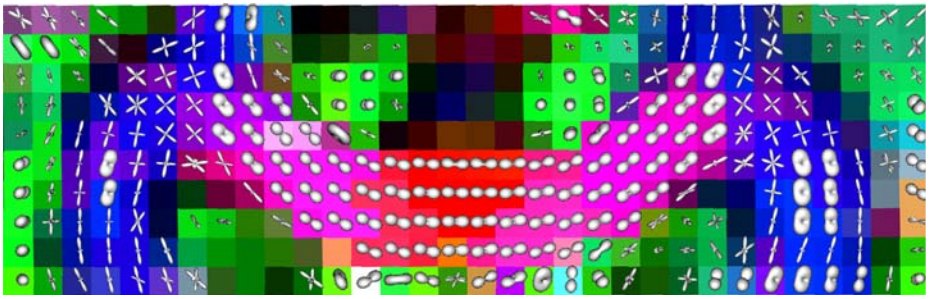


Fig. 5. Example of hybrid visualization of CSD [12] and DTI ODFs.

gain in computation compared to modeling full brain data with the same high order model. Also interaction in the visualization pipeline becomes possible, even for a full brain slice.

4 Discussion and Conclusions

Finding the correct threshold for classification, in real data is important for accurate classification, and most of the times depends on the b-value from the acquisition protocol and the angular configuration. In our study, the thresholds found in the 90 ° phantom configuration were very similar to the thresholds used to classify the real data at the same b-value. Therefore, they can be of great importance for post-processing clinically acquired data. For *in-vivo* data, a semi-automatic detection of thresholds, with feedback from the user identifying positive and negative examples can be very useful to set the classification thresholds.

There are few important messages from this work. Denoising as a preprocessing step improves the coherence of the classification areas and enhances the HARDI profiles. ADC and QBall demonstrate strong classification information, even though sometimes lack sufficient angular resolution for small crossing angle discrimination. The sharper and slightly more noisy profiles produced by DOT and its derivation (we believe this would be the case for SD techniques [6] as well) find more accurate number of maxima and are better suited for fiber tracking applications. Increasing the acquisition parameters (b-value $> 2000s/mm^2$ and NG > 80) as well as model order, does not significantly improve the classification power. In contrary, the high b-value acquisitions produce low SNR datasets that are worse for classification, and result in polluted HARDI profiles. It is even doubtful if, in practice, these higher b-value datasets improve fiber tracking.

In this work, we investigated a broad range of different anisotropy measures proposed in the literature and applied them as classification criteria for discriminating different fiber configurations within the white matter. All the measures were applied on the HARDI reconstructions and, except for CRE and NM, all measures are directly implemented on SH representation of the model that can be calculated and thresholded in real time. Some of the measures such as GA, GFA and CRE behave in similar fashion and are relatively good classification criteria. Others, such as FMI, R2, Ri, Shannon entropy, are highly dependent on the acquisition parameters and the angular configuration of the profiles, and therefore, are less reliable in clinical settings. The NM measure belongs to a different category of measures because it does not need a thresholding process for classification. However, it is very dependent on the HARDI profile and can produce many false positives in the presence of noise. A strong message that comes out of this work, is that the measures can be applied on different SPFs and still have the same classification power (especially in the case of ADC and QBall). This means that the users can use use any existing HARDI modeling technique and apply classification measures to distinguish between Gaussian and non-Gaussian profiles. If the non-Gaussian voxels are correctly classified in a first step, one can ignore all the other single fiber voxels and properly focus on the modeling and more accurate reconstruction of these voxels. Hence, as a second step, one can use a complex modeling approach, such as SD, PAS-MRI [18], etc. In clinical settings, the simplification of the data into Gaussian and non-Gaussian areas can be desired and presents a new contrast as such, even though complex structures are oversimplified as non-Gaussian. It can lead to new ways to study the white matter.

Future work will address combination of different measures for better reliability of the classified regions. Comparison of our simple and fast classification with some of the existing classification schemes as in the work of Schnell et al. [19] or learning approach such as boosting on the entire set of measures to statistically determine the discriminative strength of each feature, is addressed as future work as well. However these approaches are not interactive and real-time and the comparison should be done for validation purposes of our method only.

Nonetheless, we showed that reliable classification of Gaussian and non-Gaussian profiles can be done with some of the existing measures. The data can therefore be simplified into linear, crossing and isotropic/noise voxels. This means that more sophisticated hybrid methods, which are more time consuming can be applied only in the non-Gaussian areas, whereas linear and isotropic areas can be modeled with a simple diffusion tensor ODFs (Figure 5). This has a huge potential in the employment of the HARDI techniques in a clinical settings and enabling moderate post-processing time. Another application of the classification information can be in visualizing uncertainties in fiber tracking algorithms by attributing transparency on the unreliable fiber tracts.

Acknowledgements

We thank Alard Roebroeck from Maastricht Brain Imaging Center, Department of Cognitive Neuroscience, Faculty of Psychology, Maastricht University, The Netherlands and Pim Pullens from Brain Innovation B.V., Maastricht, The Netherlands for providing us with in-vivo datasets. This study was financially supported by the VENI program of the Netherlands Organization for Scientific Research NWO (Anna Vilanova) and by the Netherlands Organization for Scientific Research (NWO), project number 643.100.503 MFMV (Vesna Prčkovska).

References

1. Basser, P.J., Mattiello, J., Lebihan, D.: MR diffusion tensor spectroscopy and imaging. Biophys. J. 66(1), 259–267 (1994)
2. Frank, L.R.: Characterization of anisotropy in high angular resolution diffusion-weighted MRI. Magn. Reson. Med. 47(6), 1083–1099 (2002)
3. Alexander, D.C., Barker, G.J., Arridge, S.R.: Detection and modeling of non-gaussian apparent diffusion coefficient profiles in human brain data. Magn. Reson. Med. 48(2), 331–340 (2002)
4. Tuch, D.: Q-ball imaging. Magn. Reson. Med. 52, 1358–1372 (2004)
5. Özarslan, E., Shepherd, T.M., Vemuri, B.C., Blackband, S.J., Mareci, T.H.: Resolution of complex tissue microarchitecture using the diffusion orientation transform (DOT). NeuroImage 36(3), 1086–1103 (2006)
6. Jian, B., Vemuri, B.C.: A unified computational framework for deconvolution to reconstruct multiple fibers from Diffusion Weighted MRI. IEEE Transactions on Medical Imaging 26(11), 1464–1471 (2007)
7. Rao, M., Chen, Y., Vemuri, B.C., Wang, F.: Cumulative residual entropy: A new. measure of information. IEEE Transactions on Information Theory 50(6), 1220–1228 (2004)
8. Chen, Y., Guo, W., Zeng, Q., Yan, X., Rao, M., Liu, Y.: Apparent diffusion coefficient approximation and diffusion anisotropy characterization in DWI. In: Christensen, G.E., Sonka, M. (eds.) IPMI 2005. LNCS, vol. 3565, pp. 246–257. Springer, Heidelberg (2005)
9. Descoteaux, M., Angelino, E., Fitzgibbons, S., Deriche, R.: Apparent diffusion coefficients from high angular resolution diffusion imaging: Estimation and applications. Magn. Reson. in Med. 56, 395–410 (2006)
10. Özarslan, E., Vemuri, B.C., Mareci, T.H.: Generalized scalar measures for diffusion MRI using trace, variance, and entropy. Magn. Reson. Med. 53(4), 866–876 (2005)
11. Leow, A., Zhu, S., Zhan, L., McMahon, K., de Zubicaray, G., Meredith, M., Wright, M., Thompson, P.: A study of information gain in high angular resolution diffusion imaging (HARDI). In: Computational Diffusion MRI Workshop, MICCAI (2008)
12. Tournier, J.D., Calamante, F., Connelly, A.: Robust determination of the fibre orientation distribution in diffusion MRI: non-negativity constrained super-resolved spherical deconvolution. Neuroimage 35(4), 1459–1472 (2007)
13. Descoteaux, M., Angelino, E., Fitzgibbons, S., Deriche, R.: Regularized, fast and robust analytical q-ball imaging. Magn. Reson. Med. 58, 497–510 (2007)
14. Wedeen, V.J., Hagmann, P., Tseng, W.Y., Reese, T.G., Weisskoff, R.M.: Mapping complex tissue architecture with diffusion spectrum magnetic resonance imaging. Magn. Reson. Med. 54(6), 1377–1386 (2005)
15. Poupon, C., Rieul, B., Kezele, I., Perrin, M., Poupon, F., Mangin, J.F.: New diffusion phantoms dedicated to the study and validation of HARDI models. Magn. Reson. in Med. 60, 1276–1283 (2008)
16. Descoteaux, M., Wiest-Daesslé, N., Prima, S., Barillot, C., Deriche, R.: Impact of Rician Adapted Non-Local Means Filtering on HARDI. In: Metaxas, D., Axel, L., Fichtinger, G., Székely, G. (eds.) MICCAI 2008, Part II. LNCS, vol. 5242, pp. 122–130. Springer, Heidelberg (2008)

17. Prčkovska, V., Roebroeck, A.F., Pullens, W., Vilanova, A., ter Haar Romeny, B.M.: Optimal acquisition schemes in high angular resolution diffusion weighted imaging. In: Metaxas, D., Axel, L., Fichtinger, G., Székely, G. (eds.) MICCAI 2008, Part II. LNCS, vol. 5242, pp. 9–17. Springer, Heidelberg (2008)
18. Jansons, K.M., Alexander, D.: Persistent angular structure: new insights from diffusion magnetic resonance imaging data. Inverse Problems 19, 1031–1046 (2003)
19. Schnell, S., Saur, D., Kreher, B., Hennig, J., Burkhardt, H., Kiselev, V.: Fully automated classification of HARDI in vivo data using a support vector machine. NeuroImage 46(3), 642–651 (2009)

Experimental Comparison of PSNR and SSIM Metrics for Video Quality Estimation

Zoran Kotevski and Pece Mitrevski

Faculty of Technical Sciences, Ivo Lola Ribar bb, 7000 Bitola, Macedonia
{zoran.kotevski,pece.mitrevski}@uklo.edu.mk

Abstract. Since the development of digital video technology, due to the nature of digital video, the approach to video quality estimation has changed. Basically there are two types of metrics used to measure the objective quality of processed digital video: purely mathematically defined video quality metrics (DELTA, MSAD, MSE, SNR and PSNR) where the error is mathematically calculated as a difference between the original and processed pixel, and video quality metrics that have similar characteristics as the Human Visual System – HVS (SSIM, NQI, VQM) where the perceptual quality is considered in the overall video quality estimation. In this paper, an overview and experimental comparison of PSNR and SSIM metrics for video quality estimation is presented.

Keywords: Digital video, video compression, video quality metrics, HVS, SNR, PSNR, MSE, SSIM, NQI, VQM.

1 Introduction

Digital video quality estimation is concerned primarily with estimating the video quality of compressed video by means of mathematical calculations. The main goal of this compressed video quality estimation is to calculate the quality using mathematical calculations instead of estimating the quality “by hand” i.e. using larger number of human estimators [8]. If we try to define what video quality is we would end up with a conclusion that video quality is a state of perception by the Human Visual System [1]. So, this means that the best video quality estimator most definitely is the HVS. But, in real world situations, everyday availability of larger number of estimators is a huge problem and video quality metrics comes in handy. Basically there are two types of parameters for measuring the quality of processed digital video: mathematically defined metrics (DELTA, MSAD, MSE, SNR and PSNR) [5], [10], [13], [14] and metrics that have similar characteristics as the Human Visual System – HVS (SSIM, NQI, VQM) [2], [3], [4]. In order to evaluate and validate some of these video quality metrics, an experiment is conducted in which a larger number of differently processed video sequences are created and their PSNR and SSIM are measured. The results are basic charts that present these metrics dependence on the most common changes in processed video i.e. changes in brightness, contrast, hue, saturation and noise. This paper is concerned with experimental comparison of the performance of the most widely used video quality metrics – PSNR and SSIM.

2 Introduction to PSNR (Peak Signal to Noise Ratio)

The PSNR parameter is an engineering term for the ratio between the maximum possible power of a signal and the power of corrupting noise that affects the fidelity of its representation. PSNR is usually expressed in terms of the logarithmic decibel scale.

The PSNR is most commonly used as a measure of quality of reconstruction of lossy video compression coders [9], [11]. The signal in this case is the original data, and the noise is the error introduced by compression. When comparing coders it is used as an approximation to human perception of reconstruction quality. In some cases one reconstruction may appear to be closer to the original than other, even though it has a lower PSNR. Normally, higher PSNR indicates that the reconstruction is of higher quality. In ideal case the value of PSNR would be 100 dB, but in reality, in the field of image processing, typical values for PSNR are between 30 dB and 40 dB. PSNR is calculated using the mean squared error (MSE) [5], [10], [13] by the equation:

$$PSNR = 10 \cdot \log_{10}\left(\frac{255^2}{MSE}\right) \text{ [dB]} \tag{1}$$

and

$$MSE = \frac{1}{mn}\sum_{x=0}^{m-1}\sum_{y=0}^{n-1}\left[f(x,y) - g(x,y)\right]^2 , \tag{2}$$

where:

f(x,y) – is the input variable (color value of the original pixel)
g(x,y) – is the output variable (color value of the processed pixel)
m – is the number of pixels in horizontal direction
n – is the number of pixels in vertical direction

According to the mathematical equations for calculating the previously mentioned metrics (MSAD, MSE, SNR) [5], [10], [13], a conclusion can be drawn that they represent similar error values i.e. the calculated error is of the same degree. Because of this, PSNR can be considered as an unofficial representative of all the above mentioned video quality metrics. Considering its quite convenient characteristics, PSNR metric is still the most widely used metric for video quality estimation in many video processing systems, especially in video compression systems.

But, is it valid enough for us to rely on? Does it perform well enough to be taken for granted? This analysis is presented to compare PSNR to the newer SSIM metric, answer some of the questions and issues about PSNR video quality measurement and give directions to which situations it can or cannot be used.

3 Introduction to SSIM (Structural Similarity)

The Human Visual system (HVS) is highly adapted to extracting the structural information from the area of viewing. This characteristic of the HVS gives solid information that video quality metric based on extracting the structural information can provide better estimation of quality of the processed digital video [3], [7] in comparison to pure mathematical, error calculation metrics like PSNR.

The luminance of an object that is being observed is a result of the reflected light that hits its surface. Depending on the amount of light that hits the observed objects, they can appear brighter or darker, but the structure of an object is totally independent of changes in luminosity. These changes in brightness and contrast are high influential factor to the PSNR and other similar metrics and make video quality estimation deficiently accurate. Because of this, to explore the structural information of an image the influence of the luminosity should be extracted.

Structural information of an image can be defined by those characteristics that represent the structure of the objects in the scene, independently of the mean brightness and contrast [2], [3]. These measurements are based on measurement of three components: luminance comparison, contrast comparison and structure comparison. Structural similarity index is a combination of these separate components.

$$S(x,y) = f(l(x,y), c(x,y), s(x,y)) \ . \tag{3}$$

The system diagram of the structural quality assessment system is shown in Fig. 1. More details about the mathematical equations for calculating the SSIM index can be found in [3].

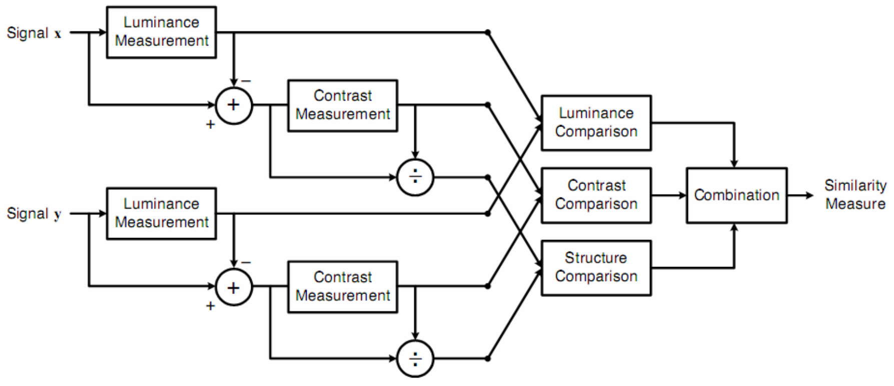


Fig. 1. Diagram of the structural similarity (SSIM) measurement system

The SSIM index can gain values from 0 to 1 where value of 1 represents maximum quality.

SSIM index is a newer quality metrics compared to PSNR and a number of cases report that SSIM performs quite better than its opponent. This experimental analysis aims to show the exact advantages and disadvantages of these two metrics.

4 Analysis of PSNR and SSIM Values in Modified Video Sequences

For the purpose of this analysis, at first three different video sequences were created. In all three videos, static picture in duration of five seconds is presented. The first video is named *Old boat* with an old boat placed by a rock. The second video is named *Sea view* and has a beautiful sea site. The third video is named *Mountains* and has a landscape of mountains and sky. The reason that three different video sequences were created is to determine if scene structure has some influence in these measurements.

Also for this purpose, approximately 300 short video sequences were produced, each with different amount of introduced changes and effects, in order to illustrate the influence of more or less visible video deformation to the performance of PSNR and SSIM metrics. The most common changes that do not highly influence the viewer's quality of experience are slight changes in brightness, contrast, hue and saturation. In other video sequences, highly visually destructive video deformation like Gaussian noise is introduced. All video sequences were created with Sony Vegas Pro v8.0c, coded in Main Concept's MPEG-2 coder, main level and profile, with average bit rate of 4 MBit/sec [15]. The measurements were performed with Elecard StreamEye Tools v2.9.1 by Elecard [12].

After the performed PSNR and SSIM calculations, the charts of their dependence on different introduced changes and effects were illustrated but, because of the limitations of this paper, in the next few pages, only some of them are presented bellow. All the charts and images of processed video sequences are publicly available at http://vq.heliohost.org.

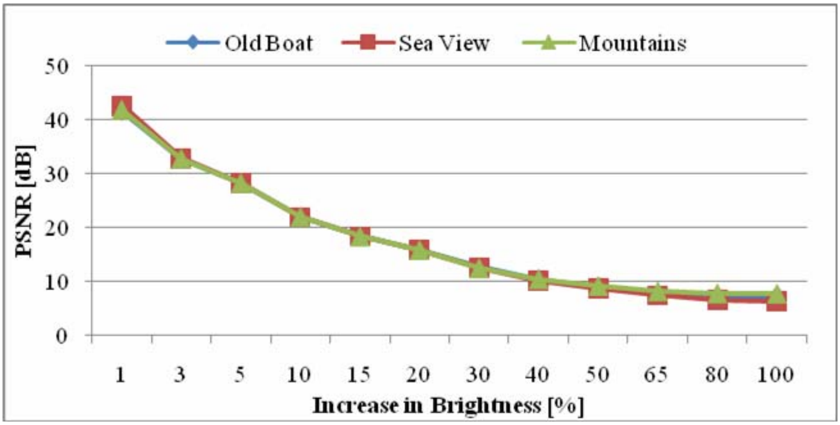


Fig. 2. PSNR decrease due to changes in brightness

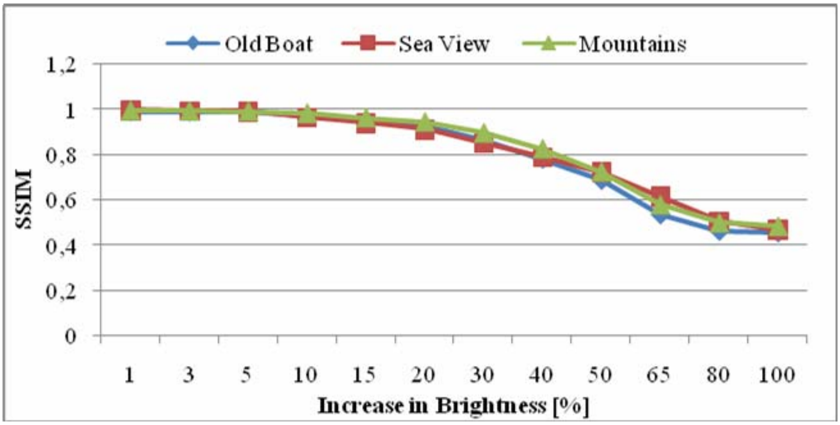


Fig. 3. SSIM decrease due to changes in brightness

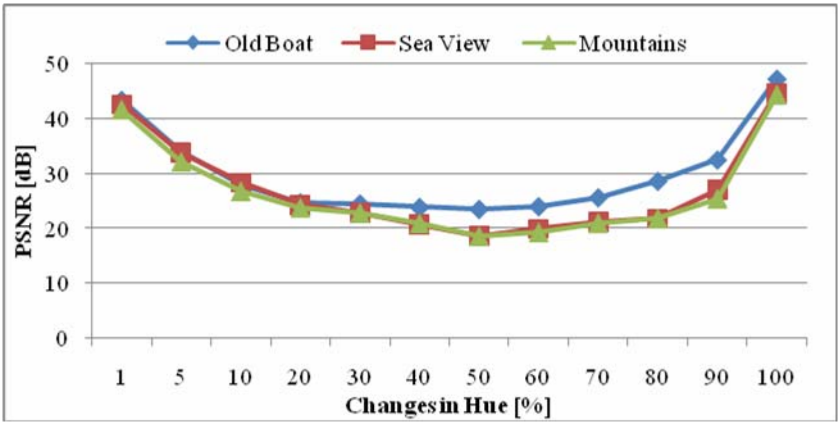


Fig. 4. PSNR changes due to changes in hue

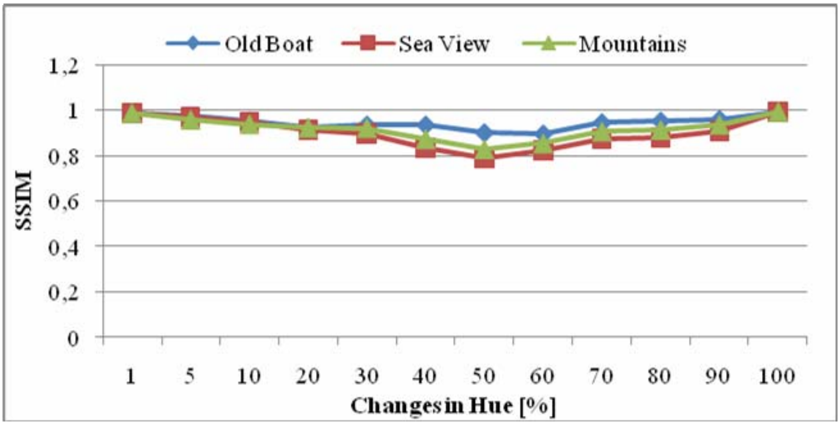


Fig. 5. SSIM changes due to changes in hue

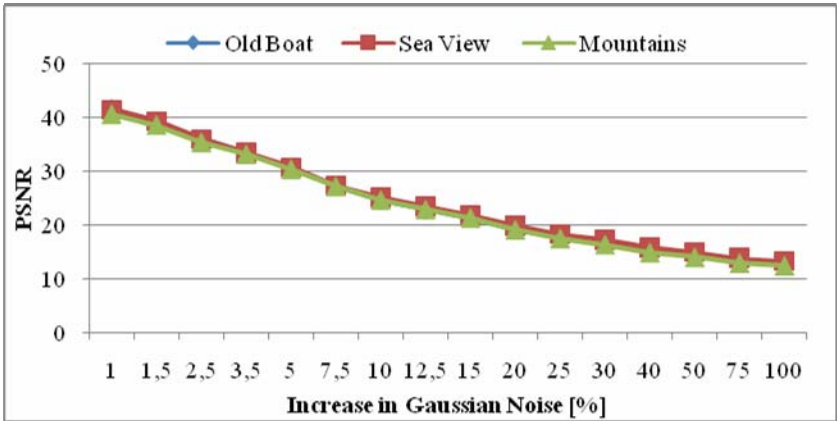


Fig. 6. PSNR decrease due to the amount of introduced Gaussian noise

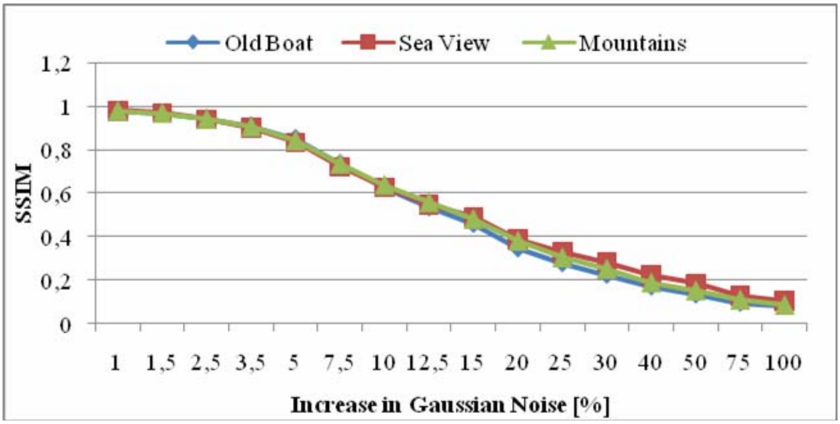


Fig. 7. SSIM decrease due to the amount of introduced Gaussian noise

After performed analysis of these charts, it can be concluded that the most drastic decreases in PSNR values are due to changes in brightness, as shown in Fig. 2, and combination of changes in brightness and contrast. The second influential factor is the introduced Gaussian noise. The changes in hue and saturation have medium effect in decreasing PSNR value. These charts clearly describe the deficiencies that PSNR metric has. Considering the performance of SSIM, the most drastic decreases in SSIM values are due to the amount of introduced Gaussian noise. Changes in brightness, contrast and hue have only mild influence to the overall quality estimation. Increase in brightness or contrast up to 25% barely influences the SSIM index, but introduction of Gaussian noise of only 10% causes quite large decrease of the SSIM index. These characteristics of SSIM, beside small deficiencies, speak of certain similarities to the HVS and present solid background for more realistic platform for quality estimation of processed digital video, compared to quality estimation using PSNR. In the next three examples, snapshots of the video sequences are presented in order to visually compare the perceived video quality with the values of the calculated PSNR and SSIM.

4.1 Example 1

In this first example, images of videos with similar PSNR are presented. PSNR metric indicates that video sequences shown in Fig. 10 and Fig. 11 are with the higher quality compared to the video sequence presented in Fig. 9. It is too obvious that in this case PSNR does not perform well. It can easily be concluded that video sequence in Fig. 9 is with noticeably better quality then the other two (Fig. 10 and Fig. 11). Some viewers would say that it looks even better than the original because of the introduced enhancements. SSIM index rates these examples much better. The image presented in Fig. 11 is obviously the one with the lowest quality. It is just a matter of question whether shifted colours as in Fig. 10 should be rated with so high quality

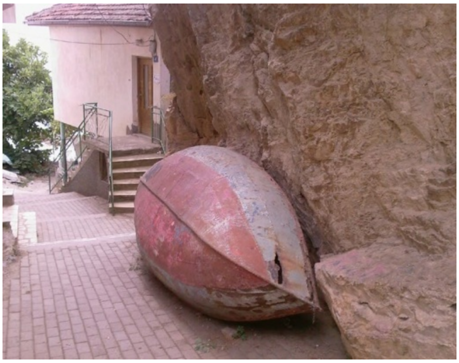


Fig. 8. The original MPEG video sequence

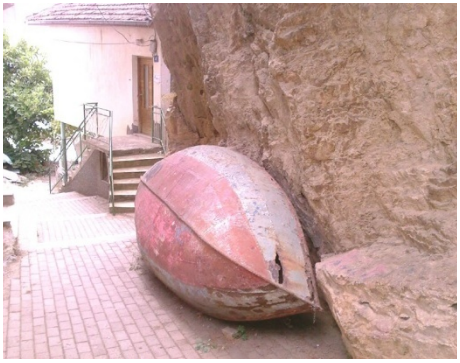


Fig. 9. Video sequence with 15% increase in brightness and contrast respectively.
PSNR=17,1768 dB, SSIM=0,9276

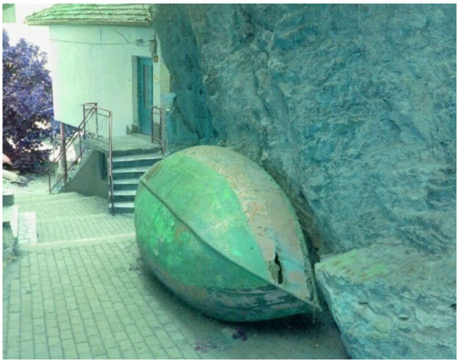


Fig. 10. Video sequence with 45% changes in hue and 45% increase in saturation.
PSNR=21,5177 dB, SSIM=0,8928

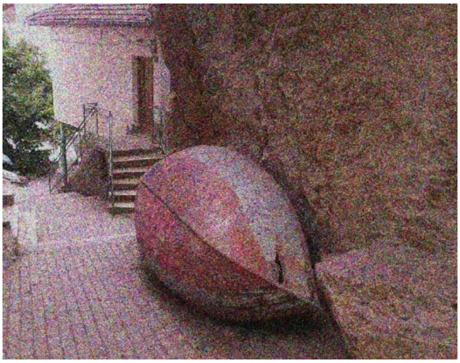


Fig. 11. Video sequence with 25% increase in Gaussian noise.
PSNR=17,6921 dB, SSIM=0,2808

index because of their visual obtrusiveness. However, the conclusion would be that SSIM performs quite better compared to PSNR in this first example.

4.2 Example 2

In the next example shown bellow, similar wrongful characteristics of PSNR metric can be concluded. Even though Fig. 15 shows noticeable noise, its PSNR value indicates highest video quality among all three measured video sequences (Fig. 13, 14, and 15). SSIM metric, once more rates them much better, giving the lowest estimated value to the image presented in Fig. 15.

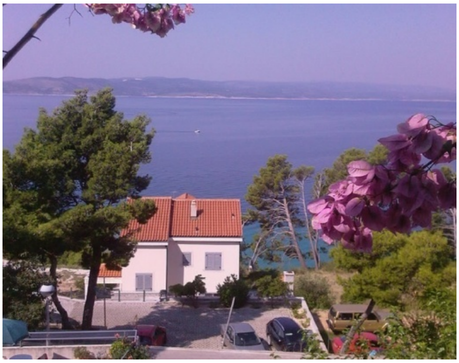


Fig. 12. The original MPEG video sequence

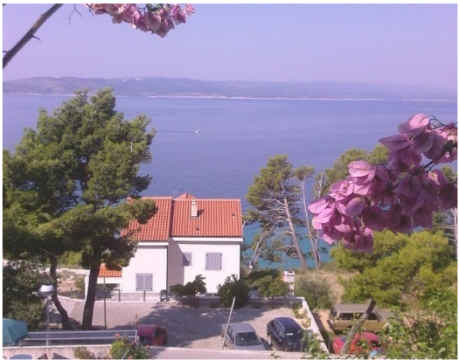


Fig. 13. Video sequence with 10% increase in brightness.
PSNR=21,9564 dB, SSIM=0,9647

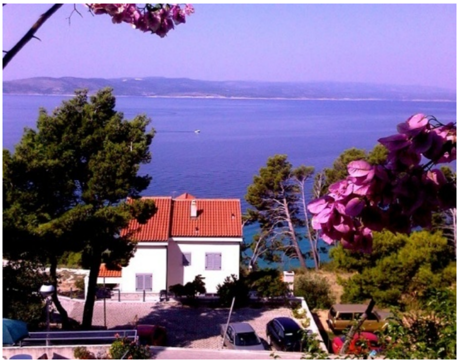


Fig. 14. Video sequence with 40% increase in contrast.
PSNR=21,5636 dB, SSIM=0,8664

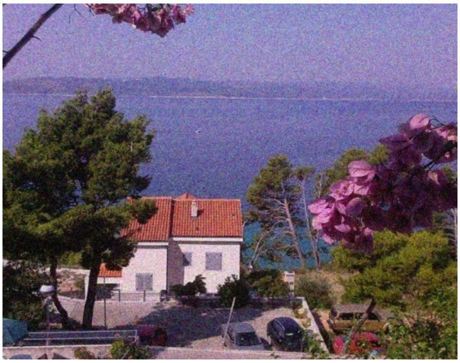


Fig. 15. Video sequence with 12.5% increase in Gaussian noise.
PSNR=23,3889 dB, SSIM=0,5484

4.3 Example 3

In the third example, opposite to previous two, images of video sequences with similar SSIM index are presented. This example shows the obvious imperfections of

SSIM also. Its low sensitivity to changes in hue or brightness results with obviously greater quality estimation error and in this particular case PSNR metric performance is quite better.

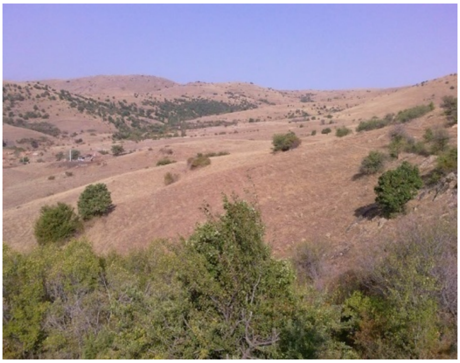


Fig. 16. The original MPEG video sequence

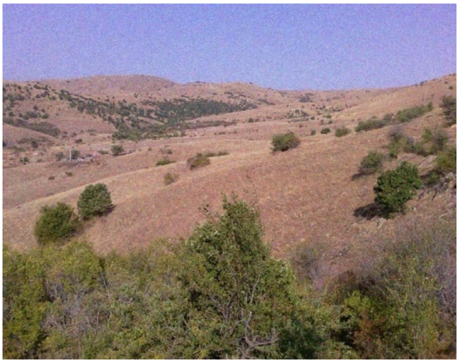


Fig. 17. Video sequence with 7,5% increase in Gaussian noise.
SSIM=0,7375, PSNR=27,1875 dB.

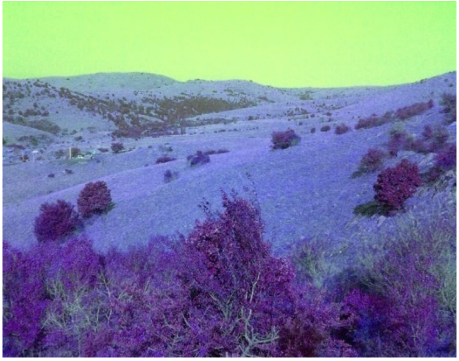


Fig. 18. Video sequence with 60% changes in hue and increase in saturation.
SSIM=0,7764, PSNR=17,1845 dB.

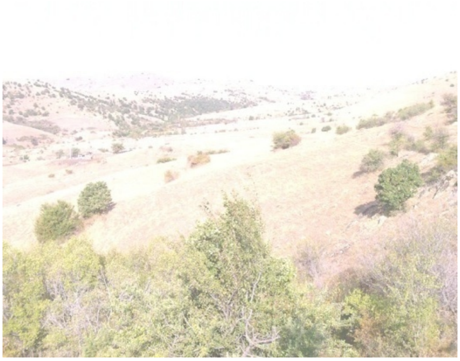


Fig. 19. Video sequence with 40% increase in brightness.
SSIM=0,8244, PSNR=10,4645 dB.

5 Conclusion

Given the examples, it can easily be concluded that PSNR metric is not valid enough to be used as objective measurement for video quality estimation. There are too many parameters that highly influence the PSNR value that are of minor visual influence to the viewer's perception of quality. Changes in brightness and contrast have high influence to the PSNR that in most cases causes decrease of performance of this metric. To be more precise, PSNR metric can be taken as valid measurement in some cases if PSNR value is greater than 35 dB. Everything below this degree of PSNR cannot be considered valid because the origin of PSNR decrease is unknown in most cases and the results given by this metric can be misleading.

On the other hand, SSIM metric has quite better performance compared to PSNR and in most cases performs very similar to the Human Visual System. But, imperfections are also present. SSIM is almost insensitive to changes in brightness, contrast and hue that when these changes are bigger SSIM values can become largely inverted. However, many examples indicate good SSIM similarity to HVS and with some small improvements in mentioned areas SSIM performance can be enhanced.

Concerning the scene structure and its influence to these measurements it can be concluded that scene composition barely influences these measurements and can be considered as non influential factor.

6 Future Research

Similar experimental comparisons are to be conducted to compare PSNR and/or SSIM to other video quality metrics (NQI, VQM, etc.) in order to propose a new basis for video quality estimation.

References

1. Kotevski, Z.: Analysis of quality and performance of MPEG-2 video compression techniques. Master Thesis, Faculty of technical sciences, Bitola, Macedonia (2007)
2. Wang, Z., Bovik, A.C.: A Universal Image Quality Index. IEEE Signal Processing Letters (March 2002)
3. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image Quality Assessment: From Error Visibility to Structural Similarity. IEEE Transactions on Image Processing (April 2004)
4. Xiao, F.: DCT Based Video Quality Evaluation. Final Project for EE392J (2000)
5. Wang, Z., Bovik, A.C.: Mean Squared Error: Love It or Leave It? IEEE Signal Processing Magazine (January 2009)
6. Wang, Z., Shang, X.: Spatial Pooling Strategies for Perceptual Image Quality Assessment. In: IEEE International Conference on Image Processing, Atlanta, GA (October 2006)
7. Wang, Z., Li, Q.: Video Quality Assessment Using a Statistical Model of Human Visual Speed Perception. Journal of the Optical Society of America (2007)
8. Richardson, I.: H.264 and MPEG-4 Video Compression. John Wiley & Sons Ltd., Chichester (2003)
9. Poynton, C.A.: A Technical Introduction to Digital Video. John Wiley & Sons Ltd., Chichester (1996)
10. Bovik, A.: Handbook of Image and Video Processing. Academic Press, London (2000)
11. Poynton, C.A.: Digital Video and HDTV – Algorithms and Interfaces. Morgan Kaufmann Publishers, San Francisco (2003)
12. Elecard, `http://www.elecard.com/`
13. Compression Project, `http://www.compression.ru/`
14. Compression Links Info, `http://www.compression-links.info/`
15. Sony Creative Software, `http://www.sonycreativesoftware.com/`

Quasigroup String Transformations and Hash Function Design

A Case Study: The NaSHA Hash Function

Aleksandra Mileva[1] and Smile Markovski[2]

[1] Faculty of Informatics, UGD, Stip, Republic of Macedonia
aleksandra.mileva@ugd.edu.mk
[2] Faculty of Natural Science, UKIM, Skopje, Republic of Macedonia
smile@ii.edu.mk

Abstract. In this paper we propose two new types of compression functions, based on quasigroup string transformations. The first type uses known quasigroup string transformations, defined elsewhere, by changing alternately the transformation direction, going forward and backward through the string. Security of this design depends of the chosen quasigroup string transformation, the order of the quasigroup and the properties satisfied by the quasigroup operations. We illustrate how this type of compression function is applied in the design of the cryptographic hash function NaSHA. The second type of compression function uses new generic quasigroup string transformation, which combine two orthogonal quasigroup operations into a single one. This, in fact, is deployment of the concept of multipermutation for perfect generation of confusion and diffusion. One implementation of this transformation is by extended Feistel network $F_{A,B,C}$which has at least two orthogonal mates as orthomorphisms: its inverse $F_{A,B,C}^{-1}$ and its square $F_{A,B,C}^{2}$.

Keywords: Compression Function, Hash Function Design, Quasigroup String Transformation, Orthogonal Quasigroups, NaSHA.

1 Introduction

Hash functions take variable-size input messages and map them into fixed-size output, known as hash result, message digest, hashcode etc. They are considered as "Swiss army knife" because of their versatile application in checking data integrity, digital signature schemes, commitment schemes, password based identification systems, digital time-stamping schemes, pseudo-random string generation, key derivation, one-time passwords etc. Almost all hash functions consist of the *compression function C* with fixed-size input and output, and the *domain extender* that, given a compression function, produces a function with variable-size input. Often, the message M is divided in blocks $M_1, M_2, ..., M_b$ with fixed size of n bits, which then are processed iteratively by the compression function. Usually, some padding rule which often contains an encoding of the length of the message is used for the last message block. The compression function C takes two inputs: a chaining variable H_i and a message block M_i. The starting chaining value is fixed to initial vector *IV*. After processing the last message block, the output from C is send to the output transformation *f* which compute the hash result *h(M)*.

The usual target of the attacks to hash functions is to find preimage, second preimage or collision. There is one group of attacks, known as *generic attacks*, that can be applied to any recent or future hash function. Generic attacks depend only of one generic parameter - the length of message digest of *n*-bits and they provide the upper security bounds to the given hash function. Time complexity of the generic random (second) preimage attack is $\mathcal{O}(2^n)$ operations, and the time complexity of the generic birthday attack is $\mathcal{O}\left(2^{\frac{n}{2}}\right)$ operations, where the "operations" correspond to the computation of the hash result for a random input. Hash function is *ideal secure* if the best attacks are the generic attacks. Second group of attacks are the *short-cut attacks*, in which for breaking the hash function, the attacker uses the flows in its design and internal structure. Hash function is said to be broken, if there is a short-cut attack faster than the best generic attack.

The most often used and standardized cryptographic hash functions are MD4, MD5, SHA-0, SHA-1 and the family of SHA-2 hash functions, which are the last standard issued by NIST. In the light of recent differential attacks [13, 14, 15] now is ongoing the NIST SHA-3 competition for a new standard for the cryptographic hash functions [12].

1.1 Our Contribution

In this paper we present two design strategies for a compression function based on quasigroup string transformations. The first design strategy was used in the cryptographic hash function NaSHA [11], which was one of the selected First Round candidates to the NIST SHA-3 competition [12]. Second design strategy is quite new in the quasigroup application in cryptography as implementation, and it deploys the concept of multipermutation for a perfect generation of confusion and diffusion [16]. It uses new quasigroup string transformation, which combine two orthogonal quasigroup operations into a single one, and it can be used without the notion of leader (needed for the first design).

1.2 Related Work

In the literature there are several compression functions based on quasigroups, but most of them are generic and not complete in a sense, that don't have any implementation or security analysis [1-5]. First serious implementation of such a hash function is one of Edon-$\mathcal{R}$ (256, 384, 512), described in [6]. The most famous example is the Edon-$\mathcal{R}$, the fastest candidate of NIST SHA-3 competition [7]. The Edon-$\mathcal{R}$ compression function uses the quasigroup reverse string transformation $\mathcal{R}$, first introduce in [8], which is special kind of E transformation [9], where the leaders are the elements of the message string, taken in revrse order. The transformation $\mathcal{R}$ is produced by fixed quasigroups of order 2^{256} and 2^{512}, isotopes of the Abelian groups $((\mathbb{Z}_2^w)^8, +_8)$, where w=32 or w=64 and $+_8$ is component wise addition on two 8-dimensional vectors in $(\mathbb{Z}_2^w)^8$.

We consider two compression function constructions with quasigroup string transformations as nonlinear building blocks. In contrast to the earlier designs, in our first type compression function, the transformation direction is changing alternately, going forward and backward through the string. Also we use different quasigroups (obtained by extended Feistel networks [10]) for each iteration of the compression function. We insert tunable parameters in the definition of the quasigroups, so they can be functions of processed message block. For a detailed discussion of the design rationale, we refer to [11], which describes the design of the cryptographic hash function NaSHA. Second design of the compression function is a new one, in the sense that it uses different orthogonal quasigroup operations for producing new orthogonal quasigroup string transformation.

2 Quasigroup String Transformations

A *quasigroup (Q, *)* is a groupoid with the property

$$(\forall\, a,b\ \in Q)(\exists!\, x,y\ \in\ Q)(a * x = b\ \ \wedge\ y * a = b). \tag{1}$$

In other words, each element will appear exactly once in each row and exactly once in each column of the multiplication table of *(Q, *)*. This means that every row and every column is a permutation of Q. To every finite quasigroup with n elements *(Q,*)*, given by its Cayley table, an equivalent combinatorial structure n by n Latin square can be associated, consisting of the matrix formed by the interior of the table.

Let $Q = \mathbb{Z}_2^n$ be an alphabet and let $*$ be a randomly chosen quasigroup operation on Q. Let denote by $Q^+ = \{\ x_1x_2 \dots\ x_t\ \mid x_i \in\ Q,\ t \geq 2\}$ the set of all finite string over Q. For a fixed letter $l \in\ Q$ called leader, the quasigroup string transformations $e_l, d_l \colon Q^+ \to Q^+$are defined in [9] as:

$$e_l(x_1 \dots x_t) = (z_1 \dots z_t) \Leftrightarrow z_j = \begin{cases} l * x_1, & j = 1 \\ z_{j-1} * x_j, & 2 \leq j \leq t \end{cases} \tag{2}$$

$$d_l(z_1 \dots z_t) = (x_1 \dots x_t) \Leftrightarrow x_j = \begin{cases} l * z_1, & j = 1 \\ z_{j-1} * z_j, & 2 \leq j \leq t \end{cases}. \tag{3}$$

Compositions of e_{l_i}or d_{l_i} transformations with fixed leaders $l_1, l_2, \dots l_s \in Q$ define new composite E and D transformations, that are permutations [9]:

$$E = e_{l_s} \circ e_{l_{s-1}} \circ e_{l_1}, \qquad D = d_{l_s} \circ d_{l_{s-1}} \circ d_{l_1}. \tag{4}$$

One can also use mix of e_l or d_l transformations as T transformation.

If we allow Q to be with group operation addition modulo 2^n, for a fixed leader $l \in Q$, the quasigroup additive string transformation $\mathcal{A}_l \colon Q^+ \to Q^+$ and the quasigroup reverse additive string transformation $\mathcal{RA}_l \colon Q^+ \to Q^+$ can be defined as [11]:

$$\mathcal{A}_l(x_1 \dots x_t) = (z_1 \dots z_t) \Leftrightarrow z_j = \begin{cases} (l + x_1) * x_1, & j = 1 \\ (z_{j-1} + x_j) * x_j, & 2 \leq j \leq t \end{cases} \tag{5}$$

$$\mathcal{RA}_l(x_1 \dots x_t) = (z_1 \dots z_t) \Leftrightarrow z_j = \begin{cases} x_j * \left(x_j + z_{j+1}\right), & 1 \leq j \leq t-1 \\ x_t * (x_t + l), & j = t \end{cases} \tag{6}$$

These transformations are not bijective mappings. Let $\mathcal{A}_{l_i}$or $\mathcal{RA}_{l_i}$ (i=1,..., s) be transformations defined by choosing fixed elements $l_1, l_2, \dots l_s \in Q$. Let m_{l_i}be any of previous $\mathcal{A}_{l_i}$or $\mathcal{RA}_{l_i}$ transformations. We can define M transformations as

$$M = m_{l_s} \circ m_{l_{s-1}} \circ m_{l_1}. \qquad (7)$$

Special kind of M transformation, so called $\mathcal{MT}$or main transformation, which is composition of $\mathcal{A}_l$ and $\mathcal{RA}_l$ transformations applied alternatively, is used in NaSHA.

Let Q be endowed with two orthogonal quasigroup operations $*_1$ and $*_2$. Then we define so called *orthogonal quasigroup string transformation* $OT: Q^+ \to Q^+$ by the following iterative procedure.

$OT(\mathrm{x}_1) = x_1$, $OT(x_1, x_2) = (x_1 *_1 x_2, x_1 *_2 x_2)$, and if $OT(x_1, x_2, \dots, x_{t-2}, x_{t-1}) = (z_1, z_2, \dots, z_{t-1})$ is defined for $t > 2$, then

$$OT(x_1, x_2, \dots, x_{t-1}, x_t) = (z_1, z_2, \dots, z_{t-1} *_1 x_t, z_{t-1} *_2 x_t),$$

where $x_i \in Q$.

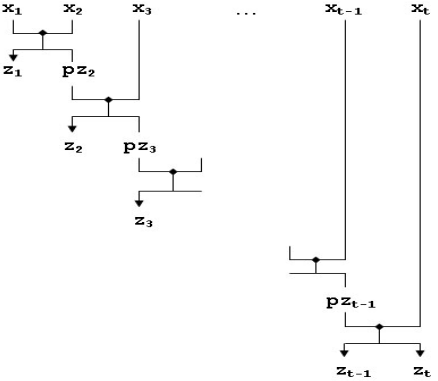


Fig. 1. Schematic representation of the orthogonal quasigroup string transformation OT

Schematic representation of $OT$is given on Fig. 1. Note that the restriction OT_n of OTon the set Q^n is a mapping $OT_n : Q^n \to Q^n$ and so $OT = OT_1 \cup OT_2 \cup OT_3 \cup \dots$, i.e., $OT$is a disjoint union of the mappings OT_n. OT_1 is the identity mapping on Q, so it is a permutation. OT_2 is a permutation of Q^2 since $(x_1 *_1 x_2, x_1 *_2 x_2) = (y_1 *_1 y_2, y_1 *_2 y_2)$ implies $(x_1, x_2) = (y_1, y_2)$ by the orthogonality of the quasigroup operations $*_1$ and $*_2$. Suppose that OT_{t-1} is a permutation for $t > 2$, and let $OT_t(x_1, x_2, \dots, x_t) = OT_t(y_1, y_2, \dots, y_t) = (z_1, z_2, \dots, z_t)$. Let $OT_{t-1}(x_1, x_2, \dots, x_{t-1}) = (u_1, u_2, \dots, u_{t-1})$ and $OT_{t-1}(y_1, y_2, \dots, y_{t-1}) = (v_1, v_2, \dots, v_{t-1})$. Then $z_1 = u_1 = v_1$, $z_2 = u_2 = v_2$, …, $z_{t-2} = u_{t-2} = v_{t-2}$ and $(z_{t-1}, z_t) = (u_{t-1} *_1 x_t, u_{t-1} *_2 x_t) = (v_{t-1} *_1 y_t, v_{t-1} *_2 y_t)$, that implies $(u_{t-1}, x_t) = (v_{t-1}, y_t)$ by orthogonality of $*_1$ and $*_2$. We have $x_t = y_t$ and $OT_{t-1}(x_1, x_2, \dots, x_{t-1}) = OT_{t-1}(y_1, y_2, \dots, \mathrm{y}_{t-1}) = (z_1, z_2, \dots, z_{t-2}, u_{t-1} = v_{t-1})$. Thus we have proved the following.

Theorem 1. The orthogonal quasigroup string transformation OT is a permutation on Q^+, and its restriction OT_n is a permutation on Q^n for each positive integer n. □

Note that if quasigroup operations are not orthogonal, the transformation defined by (8) is not necessarily a permutation.

The orthogonal quasigroup string transformation is a multipermutation. A permutation $f: Q^2 \to Q^2$, $f(a,b) = (f_1(a,b), f_2(a,b))$ is said to be a *multipermutation*, if for every $a, b \in Q$ the mappings $f_1(a,*)$, $f_1(*,b)$, $f_2(a,*)$ and $f_2(*,b)$ are permutations on Q. In general, a function $f: Q^r \to Q^n$ is said to be an *(r, n)-multipermutation* over an alphabet Q, if two different $(r+n)$-tuples of the form $(x, f(x))$ cannot collide in any r positions [21]. So, a pair of orthogonal quasigroups is a (2, 2)-multipermutation, a single quasigroup is a (2, 1)-multipermutation, and an (1,1)-multipermutation is permutation. In the light of the latest linear and differential attacks to the cryptographic primitives, the multipermutations are basic cryptographic tool for a perfect generation of diffusion, because, by changing i of the inputs at least $n - i + 1$ of the outputs will be changed [16].

Theorem 2. The restriction OT_t of an orthogonal quasigroup string transformation OT is a (t, t)-multipermutation, for each positive integer t.

Proof. OT_1 is an (1,1)- and OT_2 is a (2,2)-multipermutation. We proceed by induction, and assume that OT_k are (k, k)-multipermutations for each $k<t$.

Let $OT_t(x_1, x_2, \ldots, x_t) = (z_1, z_2, \ldots, z_t)$. We have $OT_{t-1}(x_1, x_2, \ldots, x_{t-1}) = (z_1, z_2, \ldots, z_{t-2}, u)$ and $(z_{t-1}, z_t) = (u *_1 x_t, u *_2 x_t)$. By the induction hypothesis, two different $2(t-1)$-tuples of the form $(x_1, x_2, \ldots, x_{t-1}, z_1, z_2, \ldots, z_{t-2}, u)$ cannot collide in any $t - 1$ positions. Now, suppose that two different $2t$-tuples of the form $(x_1, x_2, \ldots, x_t, z_1, z_2, \ldots, z_t)$ collide in t positions. The collision cannot happen if $t - 1$ of the positions contains some elements of the set $\{x_1, x_2, \ldots, x_{t-1}, z_1, z_2, \ldots, z_{t-2}\}$. So, the collision happens at z_{t-1}, z_t and at some $t - 2$ elements of the set $\{x_1, x_2, \ldots, x_{t-1}, z_1, z_2, \ldots, z_{t-2}\}$. From $(z_{t-1}, z_t) = (u *_1 x_t, u *_2 x_t)$, since z_{t-1} and z_t collide, there are u' and x_t' such that $(z_{t-1}, z_t) = (u' *_1 x_t', u' *_2 x_t')$. But this is a contradiction with the orthogonality of $*_1$ and $*_2$. □

Remark. Note that the orthogonal quasigroup string transformation do not use leaders.

3 Design of Compression Functions by Using Quasigroup String Transformations

We propose two new types of compression functions, based on quasigroup string transformations. The quasigroup string transformations are applied on a given string several times, with changing direction alternately, from the beginning to the end and vice versa. The new design construction is shown in Fig. 2. It consists of three layers: the linear combining layer, the quasigroup string transformation layer and the output layer. The first layer takes the message block and the chaining value as string consisting of *2b* input words S_i, where every odd indexed word is from the message and every even indexed word is from the chaining value. After that, the string is transformed with some linear transformation with high diffusion properties. The obtained string is than transformed in the second layer consisting of k consecutive quasigroup string transformations $QT^{(i)}$ that are going in different directions. The quasigroup operation for every transformation $QT^{(i)}$ is different. In this way, the influence of every bit is spreading in different directions and at the end, we obtain that its influence is spread on all words from the hash result. The output transformation can be as simple as possible, consisting of cutting first or last m bits of the transformed string. Note that a little bit more complex output function is used in NaSHA hash function [11].

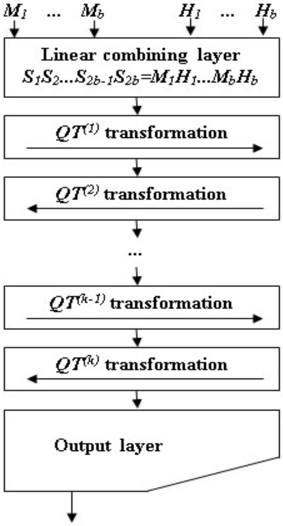

M_1 ... M_b H_1 ... H_b
Linear combining layer
$S_1S_2...S_{2b-1}S_{2b}=M_1H_1...M_bH_b$
$QT^{(1)}$ transformation
$QT^{(2)}$ transformation
...
$QT^{(k-1)}$ transformation
$QT^{(k)}$ transformation
Output layer

Fig. 2. A compression function design based on quasigroup string transformations going in different directions

The choice of quasigroup string transformations and quasigroup operations can influent the security on different ways. If we use *E, D* or *T* transformations, in fact we use permutations, so this ensures that internal state collisions can only occur in the output layer. If we use $\mathcal{A}_1$ and $\mathcal{RA}_1$ transformations, which are not permutations, the internal state collisions can occur after every applied transformation. The linear combining layer imposes relations on the input string and while they are very simple, it is hard to track them through the applied transformations, especially if they are permutations. A similar rationale can be applied for finding preimages for the compression function. To obtain more security, large shapeless quasigroups [5] of order 2^{64} and more, are preferable. If we use extended Feistel networks for quasigroup operation, we can change used quasigroup for every transformation, which additionally makes the attacker's work harder. Extended Feistel networks from the

Abelian group $(\mathbb{Z}_2^n,\oplus)$ are not a good choice for *E, D* or *T* transformations, but are a good choice for $\mathcal{A}_l$ and $\mathcal{RA}_l$ transformations because of the presence of the addition modulo operation.

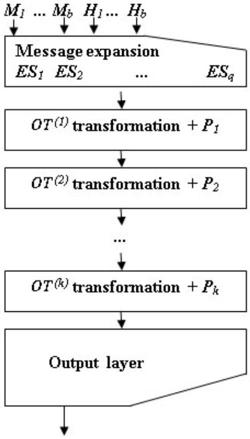

M_1 ... M_b H_1... H_b
Message expansion
ES_1 ES_2 ... ES_q
$OT^{(1)}$ transformation $+P_1$
$OT^{(2)}$ transformation $+P_2$
...
$OT^{(k)}$ transformation $+P_k$
Output layer

Fig. 3. A compression function design based on orthogonal quasigroup string transformations

The second type of compression function consists also of three layers: the message expansion layer, the orthogonal quasigroup string transformation layer and the output layer (Fig. 3). The message expansion layer takes the message block and the chaining value as in the previous construction, and then expand them to the string of *q* words ES_i. The orthogonal quasigroup string transformation layer consists of *k* application of the orthogonal quasigroup string transformations $OT^{(i)}$ (different orthogonal quasigroups in every transformation are preferable), followed by a fixed permutations. The same reasoning as for the previous case, about the collisions and the (second) preimage, holds in this case too.

4 Application of First Type of Design: The NaSHA-(*m*, 2, 6) Hash Function

NaSHA is a cryptographic hash function supporting multiple digest sizes, which uses a compression function designed according to the first design outlined in this paper. It has incorporated also the wide-pipe design of Lucks [16] and suggestions made by Coron et al [17]. For a NaSHA complete description, we refer to [11]. Here we focus on the quasigroup string transformation of NaSHA, especially of NaSHA-(*m*, 2, 6). *m* means that hash digest is *m* bits.

We use special main transformation or $\mathcal{MT}$(see Fig 4), which is defined as composition of two quasigroup string transformations (*k*=2), $\mathcal{A}_l$ and $\mathcal{RA}_l$, by shapeless quasigroups of order 2^{64}. Before applying the second transformation every word of the string is rotated to the left by half of its bits, by function ρ.

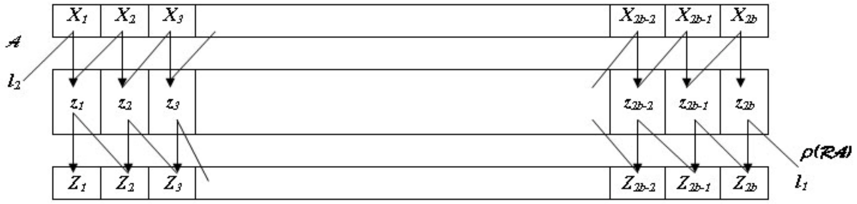


Fig. 4. Main transformation used in NaSHA-(m, 2, 6) cryptographic hash function

Quasigroup operations are implemented by extended Feistel networks from the Abelian group $(\mathbb{Z}_2^{64},\oplus)$ [10] defined as in (9) produced by fixed starting bijection of order 2^8. Different quasigroup operations for $\mathcal{A}_l$ and $\mathcal{RA}_l$ can be constructed by using different parameters for extended Feistel networks (see equations (10) and (11)), which are made to be dependable from message block.

$$F_{A,B,C}(L,R) = \left(R \oplus A, L \oplus B \oplus f_{a_1,b_1,c_1,a_2,b_2,c_2,a_3,b_3,c_3,\alpha,\beta,\gamma}(R \oplus C)\right) \quad (9)$$

$$x *_{(a_1,b_1,c_1,a_2,b_2,c_2,a_3,b_3,c_3,\alpha_1,\beta_1,\gamma_1,A_1,B_1,C_1)} y = F_{A_1,B_1,C_1}(x \oplus y) \oplus y \quad (10)$$

$$x *_{(a_1,b_1,c_1,a_2,b_2,c_2,a_3,b_3,c_3,\alpha_2,\beta_2,\gamma_2,A_2,B_2,C_2)} y = F_{A_2,B_2,C_2}(x \oplus y) \oplus y \quad (11)$$

Existing cryptanalysis [19] showed that if we do not include all state words in calculation of the tunable parameters, as was at the beginning in NaSHA for *m*=384 and *m*=512, NaSHA is susceptible on truncated differential collision attacks with unknown probability. NaSHA with modifications suggested in [20] for *m*=384 and *m*=512, is still unbroken.

5 One Implementation of OT Transformation

We propose one implementation of OT transformation with large quasigroups defined by extended Feistel network $F_{A,B,C}$ which has at least two orthogonal mates as orthomorphisms: its inverse $F_{A,B,C}^{-1}$ and $F_{A,B,C}^{2}$. Proof of this is given in Appendix A. In general, the orthomorphisms $F_{A,B,C}^{-1}$ and $F_{A,B,C}^{2}$ are not orthogonal.

So, we can use the extended Feistel networks $F_{A,B,C}$ and $F_{A,B,C}^{2}$ (or $F_{A,B,C}^{-1}$) for creating two orthogonal quasigroup operations, needed an OT transformation to be defined.

6 Conclusion

In this paper we have showed how quasigroup string transformations were applied successfully in the design of the NaSHA cryptographic hash function. We also demonstrated how multipermutations via orthogonal quasigroups can be applied in designing cryptographic hash functions. We propose one possible implementation of OT transformation with large quasigroups.

References

1. Markovski, S., Gligoroski, D., Andova, S.: Using Quasigroups for one-one Secure Encoding. In: Proceedings of VIII Conference on Logic and Computer Science, LIRA 1997, Novi Sad, pp. 157–162 (1997)
2. Dvorský, J., Ochodková, E., Snášel, V.: Hash Function based on Large Quasigroups. In: Proceedings of Velikonocni kriptologie, Brno, pp. 1–9 (2002)
3. Snášel, V., Abraham, A., Dvorský, J., Krömer, P., Platoš, J.: Hash Function based on Large Quasigroups. In: Allen, G., Nabrzyski, J., Seidel, E., van Albada, G.D., Dongarra, J., Sloot, P.M.A. (eds.) ICCS 2009. LNCS, vol. 5544, pp. 521–529. Springer, Heidelberg (2009)
4. Markovski, S., Gligoroski, D., Bakeva, V.: On Infinite Class of Strongly Collision Resistant Hash Functions "Edon-F" with Variable Length of Output. In: Proceedings of 1st Conference on Discrete Mathematics and Informatics for Industry, Thessaloniki, pp. 302–308 (2003)
5. Gligoroski, D., Markovski, S., Kocarev, L.: Edon-R, an Infinite Family of Cryptographic Hash Functions. In: The Second NIST Cryptographic Hash Workshop, UCSB, Santa Barbara, pp. 275–285 (2006)
6. Gligoroski, D., Knapskog, S.J.: Edon-R (256, 384, 512) - an Efficient Implementation of Edon-R Family of Cryptographic Hash Functions. Cryptology ePrint Archive, Report 2007/154 (2007)
7. Gligoroski, D., Ødegård, R.S., Mihova, M., Knapskog, S.J., Kocarev, L., Drápal, A., Klima, V.: Cryptographic Hash Function Edon-R. Submission to NIST SHA-3 competition (2008)
8. Gligoroski, D.: Candidate one-way Functions and one-way Permutations based on quasigroup String Transformations. Cryptology ePrint Archive, Report 2005, 352 (2005)
9. Markovski, S., Gligoroski, D., Bakeva, V.: Quasigroup String Processing – Part I. Contributions, Sec. Math. Tech. Sci., MANU, XX, 1-2, 13–28 (1999)
10. Markovski, S., Mileva, A.: Generating huge quasigroups from small non-linear bijections via extended Feistel network. Quasigroups and Related Systems 17, 91–106 (2009)
11. Markovski, S., Mileva, A.: NaSHA. Submission to NIST SHA-3 competition (2008)
12. National Institute of Standards and Technology: Announcing Request for Candidate Algorithm Nominations for a New Cryptographic Hash Algorithm (SHA-3) Family. Federal Register 72(212), 62212–62220 (November 2007)
13. Wang, X., Yu, H.: How to Break MD5 and Other Hash Functions. In: Cramer, R. (ed.) EUROCRYPT 2005. LNCS, vol. 3494, pp. 19–35. Springer, Heidelberg (2005)
14. Wang, X., Yu, H., Yin, L.: Efficient Collision Search Attacks on SHA-0. In: Shoup, V. (ed.) CRYPTO 2005. LNCS, vol. 3621, pp. 1–16. Springer, Heidelberg (2005)
15. Wang, X., Yin, L., Yu, H.: Finding Collisions in the Full SHA-1. In: Shoup, V. (ed.) CRYPTO 2005. LNCS, vol. 3621, pp. 17–36. Springer, Heidelberg (2005)
16. Schnorr, C.P., Vaudenay, S.: Black Box Cryptanalysis of Hash Networks Based on Multipermutations. In: De Santis, A. (ed.) EUROCRYPT 1994. LNCS, vol. 950, pp. 47–57. Springer, Heidelberg (1995)
17. Lucks, S.: A Failure-Friendly Design Principle for Hash Functions. In: Roy, B. (ed.) ASIACRYPT 2005. LNCS, vol. 3788, pp. 474–494. Springer, Heidelberg (2005)
18. Coron, J.-S., Dodis, Y., Malinaud, C., Puniya, P.: Merkle-damgård revisited: How to construct a hash function. In: Shoup, V. (ed.) CRYPTO 2005. LNCS, vol. 3621, pp. 430–448. Springer, Heidelberg (2005)

19. Ji, L., Liangyu, X., Xu, G.: Collision attacks on NaSHA-512. Cryptology ePrint Archive, Report 2008/519 (2008)
20. Markovski, S., Mileva, A.: NaSHA. In: First SHA-3 Candidate Conference (2008), `http://csrc.nist.gov/groups/ST/hash/sha-3/Round1/Feb2009/documents/NaSHAforweb.pdf`
21. Vaudenay, S.: On the Need for Multipermutations: Cryptanalysis of MD4 and SAFER. In: Preneel, B. (ed.) FSE 1994. LNCS, vol. 1008, pp. 286–297. Springer, Heidelberg (1995)

Appendix A

With each orthomorphism θ one can associate a quasigroup $(Q,*_\theta)$ defined as $x *_\theta y = x + \theta(y)$. Two orthomorphisms θ_1 and θ_2 are *orthogonal* if they produce orthogonal quasigroups $\left(Q,*_{\theta_1}\right)$ and $\left(Q,*_{\theta_2}\right)$. This is fulfilled if and only if the mapping $\Phi: x \to \theta_1(x) - \theta_2(x)$ is a permutation on Q.

Proposition 3. Let $F_{A,B,C}: Q^2 \to Q^2$ be an extended Feistel network of Abelian group $(\mathrm{Q}^2, +)$ created by a bijection $\mathrm{f}: \mathrm{Q} \to \mathrm{Q}$. $F_{A,B,C}$ and $F_{A,B,C}^2$ are orthogonal orthomorphisms.

Proof. Let conditions of the theorem are fulfilled. Let $\Phi = F_{A,B,C}^2 - F_{A,B,C}$. Then for every $l,\ r \in Q$, we have

$$\Phi(l,r) = \begin{pmatrix} l - r + B + f(r + C), \\ r - l + A + f\big(l + B + C + f(r + C)\big) - f(r + C) \end{pmatrix}$$

Define the function $\Omega: Q^2 \to Q^2$ by

$$\Omega(l,r) = \begin{pmatrix} -f(f^{-1}(l + r - A - B) - l) + f^{-1}(l + r - A - B) - B - C, \\ f^{-1}(l + r - A - B) - l - C \end{pmatrix}$$

It can be checked that $\Omega \circ \Phi = \Phi \circ \Omega = \mathrm{I}$, i.e., Φ and $\Omega = \Phi^{-1}$ are bijections.

Performances of Error-Correcting Codes Based on Quasigroups

Aleksandra Popovska-Mitrovikj, Smile Markovski, and Verica Bakeva

University “Ss Cyril and Methodius” in Skopje,
Faculty of Natural Sciences and Mathematics,
Institute of Informatics, P.O. Box 162, Republic of Macedonia
aleksandrap@ii.edu.mk, smile@ii.edu.mk, verica@ii.edu.mk

Abstract. In this paper we examine some performances of error-correcting codes based on quasigroup transformations proposed elsewhere. In these error-correcting codes, there exists a correlation between any two bits of a codeword. Also, these codes are nonlinear and almost random. We give simulation results of packet-error and bit-error probability for binary symmetric channel and for several parameters of these codes. From these simulation results we can conclude that the performances of these codes depend on the used quasigroup, the length of the initial key and the way of introducing the redundant information.

Keywords: error-correcting code, random code, packet-error probability, bit-error probability, quasigroup, quasigroup transformation.

1 Introduction

In the seminal work of Claude Shannon *A Mathematical Theory of Communication* [1], error correcting codes were introduced. However, writing down an explicit and practical encoder and decoder that are as good as proved in [1] is still an unsolved problem. Several error correcting codes were proposed since then. The performances of some of these codes are extremely close to the Shannon limit. These codes are based on the following philosophy: constrained random code ensembles, described by some fixed parameters plus randomness, decoded using iterative algorithms or message passing decoders. In the paper [4], a class of codes based on quasigroups is proposed. For these codes, similar to recursive convolution codes, the correlation exists between any two bits of a codeword, which can have infinite length, theoretically. However, in contrast to convolution codes, these codes are nonlinear and almost random.

Here we consider the codes defined in [4] and our task is to investigate the influence of the code parameters to the code performances. Since these codes are designed using quasigroup string transformations and by introducing redundancy, a special attention was taken how the chosen quasigroups, the way of application of quasigroup transformations and the pattern of redundancy affect the codes. For that aim, several experiments were done and the results show a way for improving the effectiveness of the decoding algorithm of the codes.

The Section 2 contains the definition of quasigroup transformations and definition of TASC (Totally Asynchronous Stream Ciphers), used for codes definition. Description of the codes, i.e., the algorithms for coding and for decoding, is given in Section 3. The experimental results are presented in Section 4, which is the section with the main results in this paper.

2 Quasigroup Transformation and TASC

A quasigroup $(Q, *)$ is a groupoid, i.e., a set Q with a binary operation $* : Q^2 \rightarrow Q$, such that for all $u, v \in Q$, there exist unique $x, y \in Q$, satisfying the equalities $u * x = v$ and $y * u = v$. In the sequel we assume that the set Q is a finite set. The cardinality of this set, $|Q|$, is called an order of the quasigroup. The main body of the multiplication table of a quasigroup is a Latin square over the set Q.

Given a quasigroup $(Q, *)$ a new operation "\", called a parastrophe, can be derived from the operation $*$ as follows:

$$x * y = z \quad \Leftrightarrow \quad y = x \setminus z.$$

Then the algebra $(Q, *, \setminus)$ satisfies the identities: $x \setminus (x * y) = y$ and $x * (x \setminus y) = y$, and $(Q, \setminus)$ is also a quasigroup.

Quasigroup string transformations are defined on a finite set Q (i.e., an alphabet Q) endowed with a quasigroup operation $*$, and they are mappings from Q^+ to Q^+, where Q^+ is the set of all nonempty words on Q. Note that $Q^+ = Q \cup Q^2 \cup Q^3 \cup \ldots$. Here, we use two types of quasigroup transformations as explained below.

Let $l \in Q$ be a fixed element, called a leader. For every a_i, $b_i \in Q$, e- and d-transformations are defined as follows.

$$e_l(a_1a_2 \ldots a_n) = b_1\, b_2 \ldots b_n \Leftrightarrow b_{i+1} = b_i * a_{i+1},$$
$$d_l(a_1a_2 \ldots a_n) = b_1\, b_2 \ldots b_n \Leftrightarrow b_{i+1} = a_i \setminus a_{i+1},$$

for each $i = 0, 1, \ldots, n - 1$, where $b_0 = a_0 = l$. By using the identities $x \setminus (x * y) = y$ and $x * (x \setminus y) = y$, we have that $d_l(e_l(a_1a_2 \ldots a_n)) = a_1a_2 \ldots a_n$ and $e_l(d_l(a_1a_2 \ldots a_n)) = a_1a_2 \ldots a_n$. This means that e_l and d_l are permutations on Q^n, mutually inverse. In the code design compositions of e_l and d_l are used.

Theorem 1. [2] Consider an arbitrary string $\alpha = a_1a_2 \ldots a_n$ where $a_i \in Q_i$, and let β be obtained after k applications of an e-transformation. If n is large enough integer then, for each $1 \leq t \leq k$, the distribution of substrings of β of length t is uniform. (We note that for $t > k$ the distribution of substrings of β of length t may not be uniform.)

The concept of TASC was introduced recently in [3]. That cryptographic concept is the corner stone for the new algorithm for error correction. Here we use a way of implementation of TASC by quasigroup string transformations. We take the alphabet $Q = \{0, 1, \ldots, 9, a, b, c, d, e, f\}$, whose elements are 4-bit words. The elements of Q^+ (stream messages) will be denoted by $M = m_1m_2 \ldots m_l$, $m_i \in Q$. Similarly by $C = c_1c_2 \ldots c_l$, $c_i \in Q$ will be denoted the encrypted stream. The empty message will be denoted by λ. In the sequel, by $k \in Q^n$ will be denoted the key of the encryption, where n is the length of the key. A TASC is a bijection $T : Q_n \times Q^+ \rightarrow Q_n \times Q^+$ with the property that the keystream is generated as a function of the intermediate key and a fixed number of previous ciphertext digits. So, for every message $M = m_1m_2 \ldots m_l$, and for every initial key $k_0 \in Q$, the mapping $T(k_0, M) = (k_1, C)$ is defined by the following encryption process:

$$k_{i+1} = f(k_i, m_i), \quad c_i = h(k_i, m_i),$$

where k_0 is the initial state of the key, f is the key next-state function, and h is the output function which nonlinearly combines the key and plaintext m_i to produce ciphertext c_i. The main characteristic of TASC is that the error propagation is unbounded and propagates until the end of the stream.

However, by adding some redundant information in the stream, the correction of some errors can be done. That is in fact the main idea behind TASC Error Correction.

3 Code Description

Code design uses the alphabet $Q = \{0, 1, \ldots, 9, a, b, c, d, e, f\}$ of nibbles and a quasigroup operation $*$ on Q, together with its parastrophe $\backslash$.

3.1 Description of Coding

Let $M = m_1m_2 \ldots m_r$ be a block of N_{block} bits, where m_i is a nibble (4-bit letter); hence, $N_{block} = 4r$. We first add redundancy as zero bits and produce block $L = L^{(1)} L^{(2)} \ldots L^{(s)} = L_1L_2 \ldots L_m$ of N bits, where $L^{(i)}$ are 4-nibble words, L_i are nibbles, so $m = 4s$, $N = 16s$. After erasing the redundant zeros from each $L^{(i)}$, the message L will produce the original message M. On this way we obtain an (N_{block}, N) code with rate $R = N_{block} / N$. The codeword is produced from L after applying the encryption algorithm in TASC given in Figure 1. For that aim, previously, a key $k = k_1k_2 \ldots k_n$ of length n nibbles should be chosen. The obtained codeword of M is $C = C_1C_2 \ldots C_m$, where C_i are nibbles.

3.2 Description of Decoding

After transmitting through a noise channel (for our experiments we use binary symmetric channel), the codeword C will be transformed to received message $D = D^{(1)} D^{(2)} \ldots D^{(s)} = D_1D_2 \ldots D_m$, where $D^{(i)}$ are blocks of 4 nibbles and D_j are nibbles. The decoding process consists of four steps: (*i*) procedure for generating the sets with predefined Hamming distance, (*ii*) inverse coding algorithm, (*iii*) procedure for generating decoding candidate sets and (*iv*) decoding rule.

Encryption	**Decryption**
Input: Key $k = k_1k_2\ldots k_n$ and message $L = L_1L_2\ldots L_m$ **Output**: message (codeword) $C = C_1C_2\ldots C_m$	**Input**: The pair $(a_1\, a_2\ldots\, a_s, k_1k_2\ldots k_n)$ **Output**: The pair $(c_1\, c_2\ldots\, c_s, K_1K_2\ldots K_n)$
For $j = 1$ to m $X \leftarrow L_j$; $T \leftarrow 0$; For $i = 1$ to n $X \leftarrow k_i * X$; $T \leftarrow T \oplus X$; $k_i \leftarrow X$; $k_n \leftarrow T$; **Output**: $C_j \leftarrow X$	For $i = 1$ to n $K_i \leftarrow k_i$; For $j = 0$ to $s - 1$ $X, T \leftarrow a_{j+1}$; $temp \leftarrow K_n$; For $i = n$ down to 2 $X \leftarrow temp \setminus X$; $T \leftarrow T \oplus X$; $temp \leftarrow K_{i-1}$; $K_{i-1} \leftarrow X$; $X \leftarrow temp \setminus X$; $K_n \leftarrow T$; $c_{j+1} \leftarrow X$; **Output**: $(c_1c_2\ldots\, c_s, K_1K_2\ldots K_n)$

Fig. 1. TASC algorithm for encryption and decryption

Generating sets with predefined Hamming distance – The probability that ≤ t bits in $D^{(i)}$ are not correct is

$$P(p;t) = \sum_{k=0}^{t} \binom{16}{k} p^k (1-p)^{16-k},$$

where p is probability of bit-error in binary symmetric channel. Let B_{max} be an integer such that $1 - P(p;\, B_{max}) \leq q_B$. Consider the set

$$H_i = \left\{ \alpha \mid \alpha \in Q^4, H(D^{(i)}, \alpha) \leq B_{max} \right\},$$

for $i = 1, 2, \ldots, s$, where $H(D^{(i)}\, , \alpha)$ is the Hamming distance between $D^{(i)}$ and α. Then, with probability at least $1 - q_B$ the block $C^{(i)}$ is an element of the set H_i, for $i = 1, 2, \ldots, s$. The cardinality of the sets H_i is

$$B_{checks} = 1 + \binom{16}{1} + \binom{16}{2} + \ldots + \binom{16}{B_{max}}$$

and the number B_{checks} determines the complexity of the decoding procedure: for finding the element $C^{(i)}$ in the set H_i, less than or equal to B_{checks} checks have to be made. Clearly, for efficient decoding the number of checks B_{checks} has to be reduced as much as possible.

Inverse coding algorithm – The inverse coding algorithm is the decrypting algorithm of TASC given in Figure 1.

Generating decoding candidate sets – The decoding candidate sets S_0, S_1, S_2, ...,S_s are defined iteratively. Let $S_0 = (k_1...k_n; \lambda)$, where λ is the empty sequence. Let $S_{i\text{-}1}$ be defined for $i \geq 1$. Then S_i is the set of all pairs $(\delta, w_1\, w_2...w_{16i})$ obtained by using the sets $S_{i\text{-}1}$ and H_i as follows (Here, w_j are bits.). For each $(\beta, w_1\, w_2...w_{16(i\text{-}1)})\in S_{i\text{-}1}$ and each element $\alpha\in H$, we apply the inverse coding algorithm with input (α, β). If the output is the pair (γ,δ) and if both sequences γ and $L^{(i)}$ have the redundant nibbles in the same positions, then the pair $(\delta, w_1\, w_2...w_{16(i\text{-}1)}c_1c_2...c_{16}) \equiv (\delta, w_1\, w_2...w_{16i})$ is an element of S_i.

Decoding rule – The decoding of the received codeword D is given by this rule: If the set S_s contains only one element $(d_1d_2...d_n, w_1\, w_2...w_{16s})$ then L = $w_1\, w_2...w_{16s}$. In this case, we say that we have a successful decoding. In the case when the set S_s contains more than one element, we say that the decoding of D is unsuccessful (of type more-candidate-errors). In the case $S_j = \emptyset$ for some $j \in\{1, ..., s\}$, then the process will be stopped (null-error appears). We conclude that for some $m \leq j$, $D^{(m)}$ contains more than B_{max} errors, resulting with $C_m \notin H$. Then, whenever it is possible, we may increase the value of B_{max} by 1 and repeat the decoding procedure for the block D again.

Theorem 2. [4] The block-error or packet-error probability of these codes is $q = 1 - (1 - q_B)^s$.

4 Experimental Results

In this section we present the simulation results done by choosing different parameters of the code. From the obtained results we have formulated several conclusions that can be used to improve the performances of the quasigroup codes. As we said, we take a binary symmetric channel as communication medium and we analyze packet-error and bit-error probabilities.

We have made the experiments in the following way. First, we extend input message using different patterns for redundant zero nibbles, and after that we encode the extended message and transmit it through a binary symmetric channel with probability p of bit error. For coding and decoding we use the codes described in Section 3. The outgoing message is decoded and if the decoding process completed successfully (the last set S_s of candidates for decoding has only one element), the decoded message is compared with the input message. If they differ at least one bit, then we say that an uncorrected-error appears. Then we compute the bit-error as Hamming distance between the input and the decoded message. Experiments showed that this type of package error occurs rarely.

In our experiments we also calculate bit-error of unsuccessful decoding resulting from more-candidate-error or null-error. This bit-error is calculated as follows.

When null-error appears, i.e., $S_i = \varnothing$, we take all the elements from the set $S_{i\text{-}1}$ and we find their maximal common prefix substring. If this substring has k bits and the length of the sent message is m bits ($k \le m$), then we compare this substring with the first k bits of the sent message. If they differ in s bits, then the resulting bit-error is $m{-}k{+}s$.

If a more-candidates-error appears we take all the elements from the set S_s and we find their maximal common prefix substring. The bit-error is computed as previous.

Cumulative bit-error is sum of all of the previously mentioned bit-errors.

The experimental results are presented in tables. In each table we give: theoretical probability of packet-error (PER_t), experimental probability of packet-error without errors of type more-candidates-error (PER_1), experimental probability of packet-error with all unsuccessfully decoded blocks (PER) and experimentally obtained probability of cumulative bit-error (BER). Results are presented for different values of bit-error probability p of binary symmetric channel and $B_{max} = 3$ and $B_{max} = 4$.

4.1 Influence of the Pattern on the Performances

First we present results obtained for different 6 patterns for redundant zero nibbles for (72, 288) code with rate R = ¼. In these experiments we use quasigroup given in Figure 2 and the initial key $k = 01234$.

*	0	1	2	3	4	5	6	7	8	9	a	b	c	d	e	f
0	3	c	2	5	f	7	6	1	0	b	d	e	8	4	9	a
1	0	3	9	d	8	1	7	b	6	5	2	a	c	f	e	4
2	1	0	e	c	4	5	f	9	d	3	6	7	a	8	b	2
3	6	b	f	1	9	4	e	a	3	7	8	0	2	c	d	5
4	4	5	0	7	6	b	9	3	f	2	a	8	d	e	c	1
5	f	a	1	0	e	2	4	c	7	d	3	b	5	9	8	6
6	2	f	a	3	c	8	d	0	b	e	9	4	6	1	5	7
7	e	9	c	a	1	d	8	6	5	f	b	2	4	0	7	3
8	c	7	6	2	a	f	b	5	1	0	4	9	e	d	3	8
9	b	e	4	9	d	3	1	f	8	c	5	6	7	a	2	0
a	9	4	d	8	0	6	5	7	e	1	f	3	b	2	a	c
b	7	8	5	e	2	a	3	4	c	6	0	d	f	b	1	9
c	5	2	b	6	7	9	0	e	a	8	c	f	1	3	4	d
d	a	6	8	4	3	e	c	d	2	9	1	5	0	7	f	b
e	d	1	3	f	b	0	2	8	4	a	7	c	9	5	6	e
f	8	d	7	b	5	c	a	2	9	4	e	1	3	6	0	f

\	0	1	2	3	4	5	6	7	8	9	a	b	c	d	e	f
0	8	7	2	0	d	3	6	5	c	e	f	9	1	a	b	4
1	0	5	a	1	f	9	8	6	4	2	b	7	c	3	e	d
2	1	0	f	9	4	5	a	b	d	7	c	e	3	8	2	6
3	b	3	c	8	5	f	0	9	a	4	7	1	d	e	6	2
4	2	f	9	7	0	1	4	3	b	6	a	5	e	c	d	8
5	3	2	5	a	6	c	f	8	e	d	1	b	7	9	4	0
6	7	d	0	3	b	e	c	f	5	a	2	8	4	6	9	1
7	d	4	b	f	c	8	7	e	6	1	3	a	2	5	0	9
8	9	8	3	e	a	7	2	1	f	b	4	6	0	d	c	5
9	f	6	e	5	2	a	b	c	8	3	d	0	9	4	1	7
a	4	9	d	b	1	6	5	7	3	0	e	c	f	2	8	a
b	a	e	4	6	7	2	9	0	1	f	5	d	8	b	3	c
c	6	c	1	d	e	0	3	4	9	5	8	2	a	f	7	b
d	c	a	8	4	3	b	1	d	2	9	0	f	6	7	5	e
e	5	1	6	2	8	d	e	a	7	c	9	4	b	0	f	3
f	e	b	7	c	9	4	d	2	0	8	6	3	5	1	a	f

Fig. 2. Quasigroup of order 16 and its parastrophe used in the experiments

We made experiments with the following 6 patterns:

patt.1	patt.2	patt.3	patt.4	patt.5	patt.6
1000 1000	1100 1100	1100 1100	1100 1100	1100 1000	1100 1100
1000 1000	0000 1100	1000 0000	1100 0000	0000 1100	1000 0000
1000 1000	1100 0000	1100 1000	0000 1100	1000 0000	1100 1100
1000 1000	1100 1100	1000 0000	1100 1100	1100 1000	1000 0000
1000 1000	0000 1100	1100 1100	0000 0000	0000 1100	1100 1100
1000 1000	1100 0000	1000 0000	1100 1100	1000 0000	1000 0000
1000 1000	1100 0000	1100 1000	1100 0000	1100 1000	1000 1000
1000 1000	0000 0000	1000 0000	0000 0000	0000 1100	1000 0000
1000 1000	0000 0000	0000 0000	0000 0000	1000 0000	00000 000

Results obtained with the first pattern are given in Table 1. They show that this pattern gives many unsuccessful completed decoding with more-candidates-error. But, it is important to note that the cardinality of the sets S_i in the process of decoding is small, so the decoding process takes much less time than for the other patterns.

Table 1. Experimental results for **patt.1**

$B_{max} = 3$					$B_{max} = 4$				
p	PER_t	PER_1	PER	BER	p	PER_t	PER_1	PER	BER
0.02	0.0043	0.0044	0.1186	0.0089	0.03	0.0015	0.0008	0.6057	0.0845
0.03	0.0197	0.0192	0.1329	0.0169	0.04	0.0055	0.0049	0.6107	0.0855
0.04	0.0554	0.0557	0.1720	0.0369	for $p > 0.04$ too much errors appear				
0.05	0.1188	0.1177	0.2258	0.0713					

From the results presented in Table 2 for the second pattern we can see that with this pattern we get better results than with the first one, but only for $B_{max} = 3$ in the decoding process. For $B_{max} = 4$ the decoding process often ends unsuccessfully with a more-candidates-error, although this pattern has a sufficient number of zero nibbles at the end. Also, for $B_{max} = 4$ with this pattern we received higher values of BER compared with the results for the first pattern.

Table 2. Experimental results for **patt.2**

$B_{max} = 3$					$B_{max} = 4$				
p	PER_t	PER_1	PER	BER	p	PER_t	PER_1	PER	BER
0.02	0.0043	0.0043	0.0061	0.0033	0.03	0.0015	0.0005	0.2425	0.2212
0.03	0.0197	0.0205	0.0225	0.0107	0.04	0.0055	0.0049	0.2505	0.2251
0.04	0.0554	0.0542	0.0559	0.0253	for $p > 0.04$ too much errors appear				
0.05	0.1188	0.1164	0.1183	0.0532					

Results obtained with the third pattern are given in Table 3. The probabilities of packet-error and bit-error for this pattern are much smaller than for the previous two patterns.

Table 3. Experimental results for **patt.3**

$B_{max} = 3$					$B_{max} = 4$				
p	PER_t	PER_1	PER	BER	p	PER_t	PER_1	PER	BER
0.02	0.0043	0.0039	0.0050	0.0027	0.03	0.0015	0.0017	0.0129	0.0097
0.03	0.0197	0.0181	0.0192	0.0095	0.04	0.0055	0.0055	0.0171	0.0113
0.04	0.0554	0.0523	0.0531	0.0253	0.05	0.0153	0.0218	0.0323	0.0219
0.05	0.1188	0.1196	0.1201	0.0556	0.06	0.0344	0.0378	0.0484	0.0305
for $p > 0.05$ too much errors appear					0.07	0.0665	0.0653	0.0757	0.0465
					0.08	0.1149	0.1191	0.1281	0.0766

Results for the fourth considered pattern are given in Table 4. Fourth pattern has enough zero blocks on the end and after blocks of information nibbles, but again experimentally obtained probability of packet and bit error are good only for B_{max} = 3.

Table 4. Experimental results for **patt.4**

B_{max} = 3					B_{max} = 4				
p	PER_t	PER_1	PER	BER	p	PER_t	PER_1	PER	BER
0.02	0.0043	0.0051	0.0116	0.0063	0.03	0.0015	0.0023	0.1859	0.1475
0.03	0.0197	0.0206	0.0274	0.0140	0.04	0.0055	0.0064	0.1959	0.1492
0.04	0.0554	0.0566	0.0624	0.0307	for p > 0.04 too much errors appear				
0.05	0.1188	0.1182	0.1250	0.0610					

Experimental results for PER (Table 5) obtained by the fifth pattern are similar to the results obtained by previous pattern, except the results for the probabilities of bit-error, which are better.

Table 5. Experimental results for **patt.5**

B_{max} = 3					B_{max} = 4				
p	PER_t	PER_1	PER	BER	p	PER_t	PER_1	PER	BER
0.02	0.0043	0.0052	0.0114	0.0033	0.03	0.0015	0.0006	0.2534	0.0476
0.03	0.0197	0.0194	0.0248	0.0112	0.04	0.0055	0.0050	0.2606	0.0529
0.04	0.0554	0.0582	0.0632	0.0299	for p > 0.04 too much errors appear				
0.05	0.1188	0.1210	0.1257	0.0632					

Results obtained with the sixth pattern given in Table 6 show that this pattern gives good results for B_{max} = 3 and $p < 0.05$ and B_{max} = 4 and $p < 0.07$.

Table 6. Experimental results for **patt.6**

B_{max} = 3					B_{max} = 4				
p	PER_t	PER_1	PER	BER	p	PER_t	PER_1	PER	BER
0.02	0.0043	0.0037	0.0055	0.0029	0.03	0.0015	0.0017	0.0457	0.0389
0.03	0.0197	0.0205	0.0222	0.0104	0.04	0.0055	0.0063	0.0533	0.0451
0.04	0.0554	0.0562	0.0581	0.0262	0.05	0.0153	0.0133	0.0590	0.0486
0.05	0.1188	0.1207	0.1221	0.0548	0.06	0.0344	0.0350	0.0800	0.0632
for p > 0.05 too much errors appear					0.07	0.0665	0.0637	0.1103	0.0827

From the experimental results obtained for all six proposed patterns we can conclude that the best results for the probability of packet-error and bit-error are obtained for the third pattern.

In the experiments with the first pattern, as we mentioned previously, we obtain very often unsuccessful finished decoding with more-candidates-error. But, from the cardinality of sets S_i in the process of decoding we may conclude that this pattern gives the smallest accumulation of elements in these sets. Therefore, the process of decoding is much faster than for other patterns. In unsuccessfully completed decoding with more-candidates-error we note that the last set of candidates contains between 2 and 13 elements, but the most often only 2 elements. The number of elements is 13 very rarely. Therefore, we make experiments with small changes in the pattern. In order to reduce the numbers of more-candidates-errors we add additional block of four zero nibbles at the end of the pattern. Indeed, the experimental results given in Table 7 show a large reduction of these decoding errors. For example, with the first pattern for $B_{max} = 3$ and $p = 0.02$ there are 1586 more-candidates-error, and with the new pattern for the same parameters we obtain only 14 such cases. But with this change in the pattern, rate of the code is reduced from R = ¼ = 0.25 to R = 72/304 = 0.2368 and the theoretical probabilities of packet-error obtained with the Theorem 2 are different.

Table 7. Experimental results for (72, 304) code

$B_{max} = 3$					$B_{max} = 4$				
p	PER_t	PER_1	PER	BER	p	PER_t	PER_1	PER	BER
0.02	0.0046	0.0045	0.0055	0.0022	0.03	0.0015	0.0018	0.0415	0.0065
0.03	0.0208	0.0220	0.0230	0.0117	0.04	0.0059	0.0059	0.0484	0.0091
0.04	0.0584	0.0547	0.0555	0.0282	0.05	0.0162	0.0163	0.0581	0.0135
0.05	0.1250	0.1267	0.1279	0.0638	0.06	0.0362	0.0373	0.0787	0.0244
for $p > 0.05$ too much errors appear					0.07	0.0700	0.0699	0.1114	0.0410
					0.08	0.1209	0.1199	0.1576	0.0676

Since this change on the first pattern gives better results for the probabilities of package and bit-error, we did experiments with a similar version of the first pattern with rate closer to ¼. Namely, the new pattern consists of two repetitions of the first pattern plus blocks of four zero nibbles on the end, i.e. the new pattern is

100010001000100010001000100010001000100010001000100010001000100010001000
100010001000
10001000100010001000100010001000100010001000100010001000100010001000 0000.

In this case, the length of input message is 144 bits and the length of encoded message is 592 bits. The new (144, 592) code has rate R = 144/592 = 0.24324. The obtained experimental results are given in Table 8. Some meaningful comparisons of these results with results from previous code cannot be made because these two codes have different parameters and different theoretical probability of packet-error. These modifications of pattern are actually an idea how we can get improvement in the failed decoding with more-candidates-error with some change of primary pattern. The price for this improvement is paid by larger block message and larger codewords.

Table 8. Experimental results for (144, 592) code

$B_{max} = 3$					$B_{max} = 4$				
p	PER_t	PER_1	PER	BER	p	PER_t	PER_1	PER	BER
0.02	0.0089	0.0081	0.0091	0.0043	0.03	0.0029	0.0026	0.0431	0.0041
0.03	0.0400	0.0409	0.0431	0.0214	0.04	0.0114	0.0137	0.0580	0.0105
0.04	0.1106	0.1105	0.1119	0.0567	0.05	0.0312	0.0307	0.0720	0.0187
for $p > 0.05$ too much errors appear					0.06	0.0694	0.0638	0.1059	0.0359
					0.07	0.1318	0.1305	0.1662	0.0686

4.2 The Influence of the Key Length on the Performances

Theoretical probability of packet-error given in Theorem 2 is determined under the assumption that the code is random. Therefore, in this theorem the more-candidates-errors are not provided. In Theorem 1 it is proved that if we apply t quasigroup transformations on a string, we obtain string where n-tuples of letters are uniformly distributed for $n \leq t$. In the design of these codes, the length of the key k determines how many times quasigroup transformations will be applied in forming of codeword. Therefore, longer key of the code gives "more random" code. This means that the results of experimental PER will be closer to the theoretical values for PER, i.e., the number of more-candidates-errors will be reduced. So, we made experiments with the third pattern (which give the best results) with key length 10. From the results given in Table 9 we can seen that in some experiments more-candidates-error are not appeared (in this case, in the table the values of PER_1 and PER are given bolded), and if they appear, their number is very small. We can conclude that when we use a longer key, we can obtain better results for PER with almost the same duration of the decoding process.

On the other side, the key length is not unique parameter which has influence of the PER. Namely, we made an experiment with key length 10 and the first pattern, but the number of more-candidates-errors was not smaller that the previous case with the shorter key. Hence, we can conclude that each parameter in this code design has great influence over the performance, i.e., the parameters are mutually dependent.

Table 9. Experimental results for key length 10

$B_{max} = 3$					$B_{max} = 4$				
p	PER_t	PER_1	PER	BER	p	PER_t	PER_1	PER	BER
0.02	0.0043	**0.0048**	**0.0048**	0.0021	0.03	0.0015	0.0022	0.0031	0.0014
0.03	0.0197	**0.0184**	**0.0184**	0.0085	0.04	0.0055	0.0053	0.0059	0.0034
0.04	0.0554	**0.0556**	**0.0556**	0.0263	0.05	0.0153	0.0147	0.0159	0.0093
0.05	0.1188	**0.1176**	**0.1176**	0.0549	0.06	0.0344	**0.0359**	**0.0359**	0.0224
for $p > 0.05$ too much errors appear					0.07	0.0665	0.0650	0.0666	0.0407
					0.08	0.1149	0.1125	0.1131	0.0649

4.3 The Influence of the Used Quasigroup

Since we work with finite sequences, the randomness of a sequence obtained by quasigroup transformations depends on the used quasigroup. So, we did experiments with several quasigroups, which showed that the choice of the quasigroup does not affect only on the values of PER and BER, but has an enormous influence on the speed of decoding. This confirms again that the choice of all parameters of these codes affect the performance of the code.

First we did experiments with cyclic quasigroup (i.e., the group) of order 16 and the length of the key 10. Decoding for the third pattern was too slow. So, we did experiment with the first pattern for binary symmetric channel with p = 0.02 and B_{max}=3, we received PER = 0.734087 and BER= 0.460359 (previously, PER=0.1186, BER=0.0089). While, for the modified first pattern we got PER = 0.7105 and BER= 0.45488 (previously, PER=0.0055, BER=0.0022). Hence, it is clear that the choice of quasigroup has enormous influence over the performance of the code.

After that we made experiments with quasigroup of order 16 obtained by direct product of the quasigroup of order 2. Experimental results obtained with this quasigroup are worse than the results for cyclic quasigroup. For the first pattern, p = 0.02 and B_{max} = 3 we got PER=0.99424 and BER=0.80869. The same experiment for the modified first pattern gave PER= 0.9935 and BER=0.8102.

In the paper [5] coefficient of period growth is analyzed. It represents how many times the period has grown (in average) after one application of the quasigroup transformation. At the end of this paper two examples of quasigroups of order 16 are given, one with a very high and one with a very low coefficient of period growth. We made experiments using these two quasigroups and the third pattern and the results are given in Table 10 and Table 11. From these tables it can be seen that there is small difference in the values of PER and BER, although the difference in the coefficients of period growth of these quasigroups is large. It seems that the coefficient of period growth has no big influence of the performances of the code, but this should be checked by using more quasigroups.

Table 10. Experimental results for the quasigroup with very high coefficient of period growth

B_{max} = 3					B_{max} = 4				
P	PER_t	PER_1	PER	BER	p	PER_t	PER_1	PER	BER
0.02	0.0043	0.0037	0.0045	0.0025	0.03	0.0015	0.0019	0.0166	0.0116
0.03	0.0197	0.0191	0.0202	0.0091	0.04	0.0055	0.0056	0.0203	0.0147
0.04	0.0554	0.0588	0.0599	0.0276	0.05	0.0153	0.0144	0.0281	0.0181
0.05	0.1188	0.1200	0.1213	0.0588	0.06	0.0344	0.0319	0.0425	0.0268
for p > 0.05 too much errors appear					0.07	0.0665	0.0641	0.0759	0.0468
					0.08	0.1149	0.1169	0.1300	0.0792

Table 11. Experimental results for the quasigroup with very low coefficient of period growth

B_{max} = 3					B_{max} = 4				
p	PER_t	PER_1	PER	BER	p	PER_t	PER_1	PER	BER
0.02	0.0043	0.0048	0.0061	0.0034	0.03	0.0015	0.0016	0.0147	0.0109
0.03	0.0197	0.0180	0.0189	0.0091	0.04	0.0055	0.0053	0.0209	0.0148
0.04	0.0554	0.0516	0.0525	0.0234	0.05	0.0153	0.0175	0.0281	0.0203
0.05	0.1188	0.1224	0.1236	0.0589	0.06	0.0344	0.0363	0.0478	0.0297
for p > 0.05 too much errors appear					0.07	0.0665	0.0653	0.0766	0.0456
					0.08	0.1149	0.1066	0.1175	0.0739

4.4 A Method for Decreasing the Number of Null-Errors

All previous experiments with different patterns and quasigroups allow us to see how these parameters affect on the number of more-candidates-errors. The changing of these parameters do not have great influence on the number of null-errors, since their number is given in Theorem 2 for the theoretical probability of packet-error, where it does not depend on these parameters. Unsuccessful decoding with null-error occurs when more than predicted B_{max} bit errors in some of the sub-blocks of encoded message are appeared during transmission. Therefore, it is clear that some of these errors can be eliminated if we cancel a few of iterations of the decoding process and we reprocess all of them or part of them with a larger value of B_{max}. With this procedure only part of these unsuccessful decoding message will be eliminated because we cannot know exactly in which iteration the correct sub-block does not enter in the set of candidates for decoding and exactly how much transmission errors are occurred in this sub-block B_{max} + 1, B_{max} +2 or more. Moreover, the cancellation of the iterations slows down the decoding, and the number of elements in sets S_i can become too large leading to unsuccessful decoding of type more-candidates-errors. To show this, we repeat the experiments with the third pattern on the following way. If in some iteration (for example i^{th}) of decoding we get an empty set, the two previous iterations ($(i$-$1)^{th}$ and $(i$-$2)^{th}$) are canceled. After that we reprocess the $(i$-$1)^{th}$ iteration with $B_{max} = B_{max} + 1$, and the next iterations continue with the old value of B_{max} . If an empty set appears again in the same iteration (i^{th} iteration), then we stop the process and decoding ends unsuccessfully. But if an empty set is obtained in a next iteration ($(i+1)^{th}$, $(i+2)^{th}$, ...) then the above procedure is repeated for that iteration again. In Table 12 is given the percentage of eliminated unsuccessfully completed decoding with the null-error using described modification in the decoding process. We can conclude that this modification gives much better results, although the decoding process can be slower.

Table 12. Percentage of eliminated unsuccessfully decoding with null-error

B_{max}= 3		B_{max}= 4	
p	Percentage of eliminated errors-null	p	Percentage of eliminated errors-null
0.02	16.67%	0.03	0.00%
0.03	18.33%	0.04	17.65%
0.04	15.01%	0.05	2.86%
0.05	18.30%	0.06	14.88%
		0.07	14.29%
		0.08	10.26%

5 Conclusions

We have considered a kind of random codes based on quasigroups. Our aim was to investigate some of the building components of these codes. Several experiments were performed with different types of quasigroups, different types of patterns and different lengths of the key. From the experiments we could conclude that not all quasigroups are good for these codes. Namely, the so called exponential (or fractal) quasigroups give far more effective decoding. Also, depending of the pattern, the key length can improve the effectiveness of decoding. The redundancy pattern should be carefully chosen in order to obtain more effective decoding. We noted that different parameters are mutually dependent. It is an open problem the most effective pattern to be chosen and to find out if it depends on the length of the codewords. A possible new kind of quasigroup transformations and their performances for these codes should be investigated as well.

References

1. Shannon, C.E.: A Mathematical Theory of Communication. Bell Sys. Tech. J. 27, 379–423, 623–656 (1948)
2. Markovski, S., Gligoroski, D., Bakeva, V.: Quasigroup String Processing: Part 1. Maced. Acad. of Sci. and Arts, Sc. Math. Tech. Scien. XX (1-2), 13–28 (1999)
3. Gligoroski, D., Markovski, S., Kocarev, L.j.: Totally Asynchronous Stream Ciphers + Redundancy = Cryptcoding. In: Aissi, S., Arabnia, H.R. (eds.) Proceedings of the 2007 International Conference on Security and management, SAM 2007, Las Vegas, June 25-28, pp. 446–451. CSREA Press (2007)
4. Gligoroski, D., Markovski, S., Kocarev, L.j.: Error-Correcting Codes Based on Quasigroups. In: Proceedings of 16th International Conference on Computer Communications and Networks (ICCCN 2007), August 13-16, pp. 165–172 (2007)
5. Dimitrova, V., Markovski, J.: On Quasigroup Pseudo Random Sequence Generators. In: Proceedings of the 1st Balkan Conference in Informatics, Thessaloniki, Greece, November 21-23, pp. 393–401 (2003)

On the Computational Asymmetry of the S-Boxes Present in Blue Midnight Wish Cryptographic Hash Function

Danilo Gligoroski[1] and Vlastimil Klima[2]

[1] Department of Telematics, Norwegian University of Science and Technology,
O.S.Bragstads plass 2B, N-7491 Trondheim, Norway
danilo.gligoroski@item.ntnu.no

[2] Independent cryptologist - consultant, Czech Republic
v.klima@volny.cz

Abstract. Blue Midnight Wish hash function is one of 14 candidate functions that are continuing in the Second Round of the SHA-3 competition. In its design it has several S-boxes (bijective components) that transform 32-bit or 64-bit values. Although they look similar to the S-boxes in SHA-2, they are also different.

It is well known fact that the design principles of SHA-2 family of hash functions are still kept as a classified NSA information. However, in the open literature there have been several attempts to analyze those design principles. In this paper first we give an observation on the properties of SHA-2 S-boxes and then we investigate the same properties in Blue Midnight Wish.

1 Introduction

Cryptographic hash functions are considered as the fundamental building part of the modern cryptography and information security. They are present in numerous protocols and schemes such as digital signatures, commitment schemes, password protection schemes, in algorithms for checking the data integrity, key derivation functions and cryptographic random number generators, authentication schemes and many others.

The most well known family of cryptographic hash functions is the so-called MD4 family to which belong the hash functions: MD4, MD5, SHA-0, SHA-1 and SHA-2.

MD4 and MD5 were designed by Ronald Rivest [1,2] and SHA family was designed by NSA and adopted by National Institut of Standards and Technology (NIST) as a US federal standard [3,4]. According to the time plan of the approved use of cryptographic hash functions, the SHA-2 functions are intended to replace SHA-1 in 2010 [4].

Being the most important part of the design of numerous cryptographic algorithms and schemes, cryptographic hash functions of the MD4 family in the last 15–20 years have been scrutinized by numerous cryptographers and we have witnessed several successful attacks and breakthroughs in their cryptanalysis. We can mention the cryptanalysis of den Boer and Bosselaers [5,6] in 1991 and 1993, Vaudenay [7] in 1995, Dobbertin [8] in 1996 and 1998, Chabaud and Joux [9] in 1998, Biham and Chen [10] in 2004, and Wang et al. [11,12,13,14] in 2005. Note that the fastest method for finding MD5 collisions (so called "Tunneling method") was discovered by Klima in 2006 [29] and it is able to generate collisions in several seconds on a standard PC. In short, the most well known cryptographic hash functions such as: MD4, MD5, SHA-0 and SHA-1, have succumbed to those attacks, but so far SHA-2 family remains unbroken.

Since SHA-2 was designed by NSA, the design principles behind its construction are not publicly available. However, several public papers produced by the academic cryptographic community have been devoted to the cryptanalysis of SHA-2 hash functions. Gilbert and Handschuh in 2003 have made an analysis of the SHA-2 family [15]. They proved that there exist XOR-differentials that give a 9-round local collision with probability 2^{-66}. In 2004, Hawkes, Paddon and Rose [16] improved the result and showed existence of addition-differentials of 9-round local collisions with probability of 2^{-39}. Different variants of SHA-256 have been analyzed in 2005 by Yoshida and Biryukov [17] and by Matusiewicz et al., [18]. In 2006, Mendel et al. [19], found XOR-differentials for 9-round local collisions, also with probability 2^{-39} (recently improved to the value 2^{-38} [20]). In 2008, Nikolìc and Biryukov have found collisions in 21 step reduced SHA-256, and their attack was afterwards improved by Indesteege et al., up to 24 steps [22].

Following the developments in the field of cryptographic hash functions, NIST organized two cryptographic hash workshops [23] in 2005 and 2006 respectively. As a result of those workshops, NIST decided to run a 4 year world-wide open hash competition for selection of the new cryptographic hash standard SHA-3 [24]. The requirements for the hash digest size for the new cryptographic hash functions are: 224, 256, 384 and 512 bits - the same as for the current SHA-2 standard. Out of 64 initial submissions, 51 entered the First Round [25], and 14 have been selected for the Second Round of the SHA-3 competition [26].

Blue Midnight Wish hash function is the fastest hash function among 14 candidates in the Second Round of the SHA-3 competition [27]. It has several bijective components (S-boxes) that look like the bijective components in SHA-2. In this paper we will describe some of the principles how these components were chosen showing also comparison between the similar bijective components that are present in SHA-2 functions.

The paper is organized as follows: In Section 2 we give some basic observations on the properties of the four S-boxes present in SHA-2 design, in Section 3 we analyze the S-boxes of Blue Midnight Wish according to the observed properties of SHA-2 S-boxes, and we end the paper with Conclusions and future work.

2 Observations on Some Properties of the SHA-2 S-Boxes

SHA-2 is actually a family of four hash functions with outputs of 224, 256, 384 and 512 bits, and accordingly, sometimes SHA-2 functions are denoted as

SHA-224, SHA-256, SHA-384 and SHA-512. The full description of SHA-2 family can be found in [4].

The main difference between those four functions is that SHA-224 and SHA-256 are defined by operations performed on 32-bit variables, while SHA-384 and SHA-512 are defined by operations performed on 64-bit variables.

We give here the definitions of four S-boxes (or bijective transformations) present in the design of SHA-2 that acts either on 32 or 64 bits, while the rest of the design specifics are not important for this paper.

For SHA-224/256 those four bijective transformations are defined as:

$$\begin{array}{rl}
\Sigma_0^{256}(x) = & ROTR^2(x) \oplus ROTR^{13}(x) \oplus ROTR^{22}(x) \\
\Sigma_1^{256}(x) = & ROTR^6(x) \oplus ROTR^{11}(x) \oplus ROTR^{25}(x) \\
\sigma_0^{256}(x) = & ROTR^7(x) \oplus ROTR^{18}(x) \oplus \ SHR^3(x) \\
\sigma_1^{256}(x) = & ROTR^{17}(x) \oplus ROTR^{19}(x) \oplus \ SHR^{10}(x)
\end{array} \tag{1}$$

where $ROTR^n(x)$ means rotation of the 32-bit variable x to the right for n positions and $SHR^n(x)$ means shifting of the 32-bit variable x to the right for n positions.

For SHA-384/512 the four bijective transformations are defined as:

$$\begin{array}{rl}
\Sigma_0^{512}(x) = & ROTR^{28}(x) \oplus ROTR^{34}(x) \oplus ROTR^{39}(x) \\
\Sigma_1^{512}(x) = & ROTR^{14}(x) \oplus ROTR^{18}(x) \oplus ROTR^{41}(x) \\
\sigma_0^{512}(x) = & ROTR^1(x) \oplus ROTR^8(x) \oplus \ SHR^7(x) \\
\sigma_1^{512}(x) = & ROTR^{19}(x) \oplus ROTR^{61}(x) \oplus \ SHR^6(x)
\end{array} \tag{2}$$

where $ROTR^n(x)$ means rotation of the 64-bit variable x to the right for n positions and $SHR^n(x)$ means shifting of the 64-bit variable x to the right for n positions.

Previously, an interest to analyze the S-boxes in SHA-2 was described in the work of Matusiewicz et al., [18] where they noted:

- *"The substitution boxes Σ_0 and Σ_1 constitute the essential part of the hash function and fulfil two tasks: they add bit diffusion and destroy the ADD-linearity of the function."*
- *"σ_0 and σ_1 have both the property to increase the Hamming weight of low-weight inputs. This increase is upper bounded by a factor of 3. The average increase of Hamming weight for low-weight inputs is even higher if three rotations are used instead of two rotations and one bit-shift. However, a reason for this bit-shift is given by the next observation."*
- *"In contrast to all other members of the MD4-family including SHA-1, rotating expanded message words to get new expanded message words is not possible anymore (even in the XOR-linearised case). This is due to the bit-shift being used in σ_0 and σ_1."*

In what follows we will give another observation for the used S-boxes in SHA-2. For that purpose let us recall the following simple fact:

Corollary 1. *The relations expressed in equations (1) and (2) can be expressed in a matrix-vector form:*

$$\begin{array}{rcl} \Sigma_0^{256}(x) &=& \mathbf{\Sigma}_0^{256} \cdot x \\ \Sigma_1^{256}(x) &=& \mathbf{\Sigma}_1^{256} \cdot x \\ \sigma_0^{256}(x) &=& \mathbf{s}_0^{256} \cdot x \\ \sigma_1^{256}(x) &=& \mathbf{s}_1^{256} \cdot x \end{array} \tag{3}$$

$$\begin{array}{rcl} \Sigma_0^{512}(x) &=& \mathbf{\Sigma}_0^{512} \cdot x \\ \Sigma_1^{512}(x) &=& \mathbf{\Sigma}_1^{512} \cdot x \\ \sigma_0^{512}(x) &=& \mathbf{s}_0^{512} \cdot x \\ \sigma_1^{512}(x) &=& \mathbf{s}_1^{512} \cdot x \end{array} \tag{4}$$

where $\mathbf{\Sigma}_0^{256}$, $\mathbf{\Sigma}_1^{256}$, $\mathbf{s}_0^{256}$ *and* $\mathbf{s}_1^{256}$ *are* 32×32 *nonsingular matrices in* $GF(2)$, *and where* $\mathbf{\Sigma}_0^{512}$, $\mathbf{\Sigma}_1^{512}$, $\mathbf{s}_0^{512}$ *and* $\mathbf{s}_1^{512}$ *are* 64×64 *nonsingular matrices in* $GF(2)$ *and the vector* x *is 32 dimensional in equation (3) or is 64 dimensional in equation (4).* □

For the properties that we have observed on SHA-2 S-boxes we need the following Lemma:

Lemma 1. *Every nonsingular matrix* $\mathbf{S}$ *of order* $n \times n$ *in* $GF(2)$, *is also nonsingular in the ring* $\mathbb{Z}_{2^n}(+,*)$ *where the operation "+" is addition modulo* 2^n *and the operation "*" is multiplication modulo* 2^n. □

We have used Lemma 1 and interpreted the matrices $\mathbf{\Sigma}_0^{256}$, $\mathbf{\Sigma}_1^{256}$, $\mathbf{s}_0^{256}$ and $\mathbf{s}_1^{256}$ in the ring $\mathbb{Z}_{2^{32}}(+,*)$, counting the number of different elements present in their inverses: $\left(\mathbf{\Sigma}_0^{256}\right)^{-1}$, $\left(\mathbf{\Sigma}_1^{256}\right)^{-1}$, $\left(\mathbf{s}_0^{256}\right)^{-1}$ and $\left(\mathbf{s}_1^{256}\right)^{-1}$.

We did that too for $\left(\mathbf{\Sigma}_0^{512}\right)^{-1}$, $\left(\mathbf{\Sigma}_1^{512}\right)^{-1}$, $\left(\mathbf{s}_0^{512}\right)^{-1}$ and $\left(\mathbf{s}_1^{512}\right)^{-1}$.

Before presenting the results of our analysis of S-boxes used in SHA-2, let us formalize our observations by the following Definition:

Definition 1. *For every nonsingular matrix* $\mathbf{S}$ *of order* $n \times n$ *in* $GF(2)$, *let us denote by* $C(\mathbf{S^{-1}})$ *the number of different elements present in the inverse matrix* $\mathbf{S^{-1}}$ *when the inverse is taken in the ring* $\mathbb{Z}_{2^n}(+,*)$.

For example let us take $n = 16$ and let $\mathbf{S}$ be the following matrix:

$$\mathbf{S} = \begin{pmatrix}
0&0&0&0&0&1&0&1&0&0&1&0&0&1&1&0\\
0&0&0&0&0&0&1&0&1&0&0&1&0&0&1&1\\
1&0&0&0&0&0&0&1&0&1&0&0&1&0&0&1\\
1&1&0&0&0&0&0&0&1&0&1&0&0&1&0&0\\
0&1&1&0&0&0&0&0&0&1&0&1&0&0&1&0\\
0&0&1&1&0&0&0&0&0&0&1&0&1&0&0&1\\
1&0&0&1&1&0&0&0&0&0&0&1&0&1&0&0\\
0&1&0&0&1&1&0&0&0&0&0&0&1&0&1&0\\
0&0&1&0&0&1&1&0&0&0&0&0&0&1&0&1\\
1&0&0&1&0&0&1&1&0&0&0&0&0&0&1&0\\
0&1&0&0&1&0&0&1&1&0&0&0&0&0&0&1\\
1&0&1&0&0&1&0&0&1&1&0&0&0&0&0&0\\
0&1&0&1&0&0&1&0&0&1&1&0&0&0&0&0\\
0&0&1&0&1&0&0&1&0&0&1&1&0&0&0&0\\
0&0&0&1&0&1&0&0&1&0&0&1&1&0&0&0\\
0&0&0&0&1&0&1&0&0&1&0&0&1&1&0&0
\end{pmatrix}.$$

The inverse matrix $\mathbf{S^{-1}}$ in $\mathbb{Z}_{2^{16}}(+,*)$ is the following matrix:

$$\mathbf{S^{-1}} = \begin{pmatrix}
16191 & 60910 & 46261 & 48574 & 12336 & 7710 & 53971 & 50116 & 62452 & 57055 & 19275 & 56284 & 771 & 57826 & 11565 & 15420 \\
15420 & 16191 & 60910 & 46261 & 48574 & 12336 & 7710 & 53971 & 50116 & 62452 & 57055 & 19275 & 56284 & 771 & 57826 & 11565 \\
11565 & 15420 & 16191 & 60910 & 46261 & 48574 & 12336 & 7710 & 53971 & 50116 & 62452 & 57055 & 19275 & 56284 & 771 & 57826 \\
57826 & 11565 & 15420 & 16191 & 60910 & 46261 & 48574 & 12336 & 7710 & 53971 & 50116 & 62452 & 57055 & 19275 & 56284 & 771 \\
771 & 57826 & 11565 & 15420 & 16191 & 60910 & 46261 & 48574 & 12336 & 7710 & 53971 & 50116 & 62452 & 57055 & 19275 & 56284 \\
56284 & 771 & 57826 & 11565 & 15420 & 16191 & 60910 & 46261 & 48574 & 12336 & 7710 & 53971 & 50116 & 62452 & 57055 & 19275 \\
19275 & 56284 & 771 & 57826 & 11565 & 15420 & 16191 & 60910 & 46261 & 48574 & 12336 & 7710 & 53971 & 50116 & 62452 & 57055 \\
57055 & 19275 & 56284 & 771 & 57826 & 11565 & 15420 & 16191 & 60910 & 46261 & 48574 & 12336 & 7710 & 53971 & 50116 & 62452 \\
62452 & 57055 & 19275 & 56284 & 771 & 57826 & 11565 & 15420 & 16191 & 60910 & 46261 & 48574 & 12336 & 7710 & 53971 & 50116 \\
50116 & 62452 & 57055 & 19275 & 56284 & 771 & 57826 & 11565 & 15420 & 16191 & 60910 & 46261 & 48574 & 12336 & 7710 & 53971 \\
53971 & 50116 & 62452 & 57055 & 19275 & 56284 & 771 & 57826 & 11565 & 15420 & 16191 & 60910 & 46261 & 48574 & 12336 & 7710 \\
7710 & 53971 & 50116 & 62452 & 57055 & 19275 & 56284 & 771 & 57826 & 11565 & 15420 & 16191 & 60910 & 46261 & 48574 & 12336 \\
12336 & 7710 & 53971 & 50116 & 62452 & 57055 & 19275 & 56284 & 771 & 57826 & 11565 & 15420 & 16191 & 60910 & 46261 & 48574 \\
48574 & 12336 & 7710 & 53971 & 50116 & 62452 & 57055 & 19275 & 56284 & 771 & 57826 & 11565 & 15420 & 16191 & 60910 & 46261 \\
46261 & 48574 & 12336 & 7710 & 53971 & 50116 & 62452 & 57055 & 19275 & 56284 & 771 & 57826 & 11565 & 15420 & 16191 & 60910 \\
60910 & 46261 & 48574 & 12336 & 7710 & 53971 & 50116 & 62452 & 57055 & 19275 & 56284 & 771 & 57826 & 11565 & 15420 & 16191
\end{pmatrix},$$

so $C(\mathbf{S^{-1}}) = 16$ because the matrix $\mathbf{S^{-1}}$ has these 16 different elements: {771, 7710, 11565, 12336, 15420, 16191, 19275, 46261, 48574, 50116, 53971, 56284, 57055, 57826, 60910, 62452}.

The measure $C(\mathbf{S^{-1}})$ can be seen as a concept close to the concept of one-wayness of the bijective transformations i.e. close to the concept of the computational asymmetry as defined in [28] and the references there. However, in this moment we do not have a defined strong and precise mathematical connection between our measure $C(\mathbf{S^{-1}})$ and the concept of the computational asymmetry.

By simple application of the Definition 1 we have obtained the following result:

Corollary 2.
$C\big(\mathbf{\Sigma}_0^{256^{-1}}\big) = 32$, $C\big(\mathbf{\Sigma}_1^{256^{-1}}\big) = 32$, $C\big(\mathbf{s}_0^{256^{-1}}\big) = 504$ *and* $C\big(\mathbf{s}_1^{256^{-1}}\big) = 121.$
$C\big(\mathbf{\Sigma}_0^{512^{-1}}\big) = 64$, $C\big(\mathbf{\Sigma}_1^{512^{-1}}\big) = 64$, $C\big(\mathbf{s}_0^{512^{-1}}\big) = 116$ *and* $C\big(\mathbf{s}_1^{512^{-1}}\big) = 2044.$ □

Since S-boxes in SHA-2 are obtained either by only three rotations or by two rotations and one shift to the right, we were interested to see what are the other statistical properties of the whole set of all possible S-boxes that can be obtained either by three rotations or by two rotations and one shift to the right, operating on 32 or 64 bits.

Our findings are presented in the next three Corollaries (the proofs of all of them can be done by simple exhaustive search).

Corollary 3. *If the function* $\mathbf{\Sigma} : \{0,1\}^n \rightarrow \{0,1\}^n$ *is defined as*

$$\Sigma(x) = ROTR^{r_1}(x) \oplus ROTR^{r_2}(x) \oplus ROTR^{r_3}(x) \equiv (\mathbf{\Sigma}) \cdot x \tag{5}$$

where $n = 32$ *or* $n = 64$ *and* $0 \leq r_1 < r_2 < r_3 < n$, *then* $Max(C(\mathbf{\Sigma}^{-1})) = n.$ □

Corollary 4. *If the function* $\mathbf{s} : \{0,1\}^{32} \rightarrow \{0,1\}^{32}$ *is defined as*

$$\sigma(x) = ROTR^{r_1}(x) \oplus ROTR^{r_2}(x) \oplus SHR^{r_3}(x) \equiv \mathbf{s} \cdot x \tag{6}$$

where $0 \leq r_1 < r_2 < 32$, $0 \leq r_3 < 32$, *then* $Max(C(\mathbf{s}^{-1})) = 523.$ □

Corollary 5. *If the function* $\mathbf{s} : \{0,1\}^{64} \rightarrow \{0,1\}^{64}$ *is defined as*

$$\sigma(x) = ROTR^{r_1}(x) \oplus ROTR^{r_2}(x) \oplus SHR^{r_3}(x) = \mathbf{s} \cdot x \tag{7}$$

where $0 \leq r_1 < r_2 < 32$, $0 \leq r_3 < 32$, *then* $Max(C(\mathbf{s}^{-1})) = 2079.$ □

It is noticeable that NSA designers of SHA-2 have chosen some of the S-boxes to have the maximal possible value, i.e. the values of $C\big(\mathbf{\Sigma}_0^{256^{-1}}\big)=32$, $C\big(\mathbf{\Sigma}_1^{256^{-1}}\big)=32$, $C\big(\mathbf{\Sigma}_0^{512^{-1}}\big)=64$, $C\big(\mathbf{\Sigma}_1^{512^{-1}}\big)=64$. They have also chosen two of the S-boxes with almost maximal values i.e. $C\big(\mathbf{s}_0^{256^{-1}}\big)=504$ and $C\big(\mathbf{s}_1^{512^{-1}}\big)=2044$.

For $n=32$, the total distribution of $C(\mathbf{\Sigma}^{-1})$ i.e. when $\Sigma(x)=ROTR^{r_1}(x)\oplus ROTR^{r_2}(x)\oplus ROTR^{r_3}(x)\equiv(\mathbf{\Sigma})\cdot x$ is given on Figure 1. There are in total 4960 S-boxes of type $\mathbf{\Sigma}$ and as we can see, there are just 8 possible values for $C(\mathbf{\Sigma}^{-1})$, forming the set $\{2,3,6,8,10,15,17,32\}$. The majority of those S-boxes (almost 62%) belongs to the category with 32 different elements in their inverse matrix.

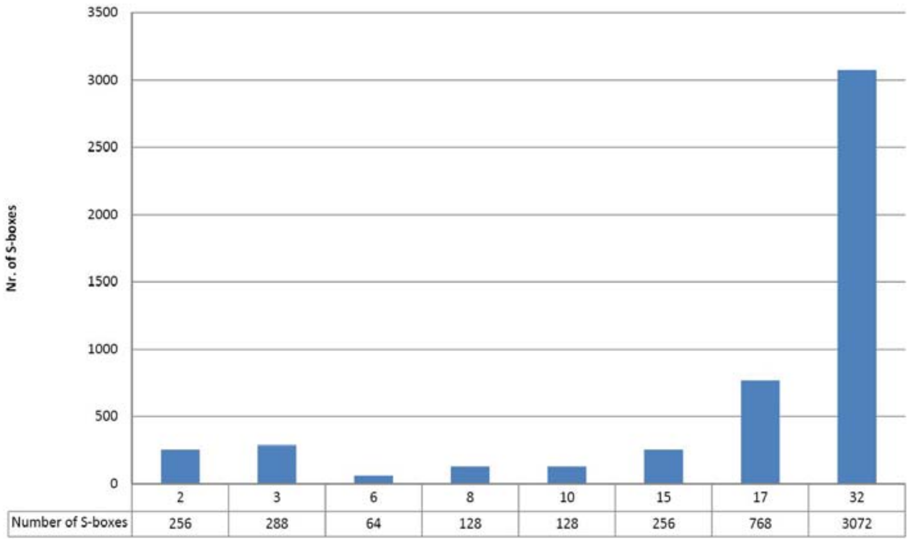
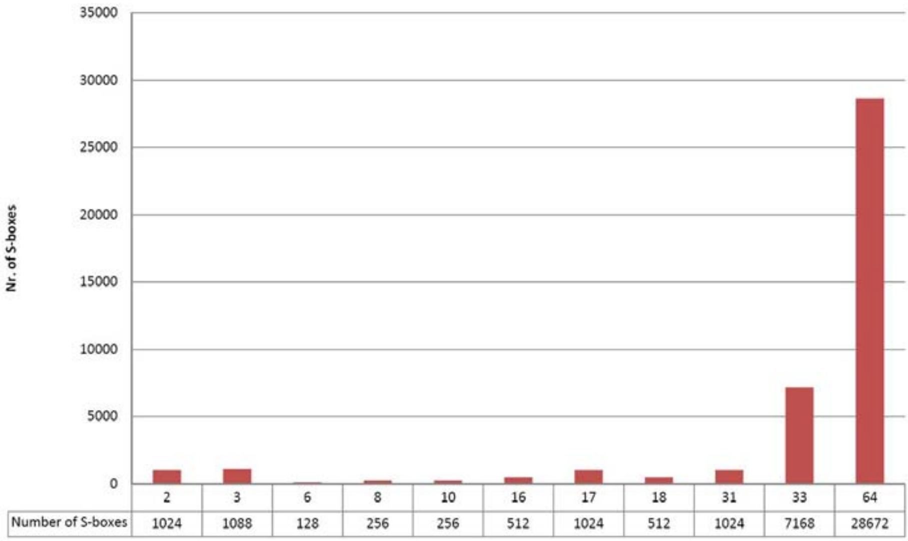


Fig. 1. A distribution of all possible values of $C(\mathbf{\Sigma}^{-1})$ for $n=32$

The distribution of $C(\mathbf{s}^{-1})$ i.e. when $\sigma(x)=ROTR^{r_1}(x)\oplus ROTR^{r_2}(x)\oplus SHR^{r_3}(x)\equiv(\mathbf{s})\cdot x$ is pretty different (and not very appropriate for graphical presentation). There are in total 489 different categories of S-boxes of type $\mathbf{s}$ according to the value of $C(\mathbf{s}^{-1})$, where minimal value is 3 and maximal value is 523.

For $n=64$, the total distribution of $C(\mathbf{\Sigma}^{-1})$ i.e. when $\Sigma(x)=ROTR^{r_1}(x)\oplus ROTR^{r_2}(x)\oplus ROTR^{r_3}(x)\equiv(\mathbf{\Sigma})\cdot x$ is given on Figure 2. There are in total 41664 S-boxes of type $\mathbf{\Sigma}$ and as we can see, there are just 11 categories of possible values for $C(\mathbf{\Sigma}^{-1})$, forming the set $\{2,3,6,8,10,16,17,18,31,33,64\}$. The majority of those S-boxes (almost 69%) belongs to the category with 64 different elements in their inverse matrix.

Similarly, the distribution of $C(\mathbf{s}^{-1})$ i.e. when $\sigma(x)=ROTR^{r_1}(x)\oplus ROTR^{r_2}(x)\oplus SHR^{r_3}(x)\equiv(\mathbf{s})\cdot x$ is pretty different. There are in total 63923



	2	3	6	8	10	16	17	18	31	33	64
Number of S-boxes	1024	1088	128	256	256	512	1024	512	1024	7168	28672

Fig. 2. A distribution of all possible values of $C(\mathbf{\Sigma}^{-1})$ for $n = 64$

S-boxes distributed in 2038 categories according to the value of $C(\mathbf{s}^{-1})$, where minimal value is 3 and maximal value is 2079.

So, as a conclusion from this analysis of SHA-2 S-boxes we can say that NSA have chosen majority of the S-boxes (6 out of 8 S-boxes) to have the property that they have maximal or close to maximal value of $C(\mathbf{\Sigma}^{-1})$ or $C(\mathbf{s}^{-1})$. Since the design principles for SHA-2 are still kept classified, we do not know is this observation just a coincidence or there is a stronger mathematical connection.

3 Properties of Blue Midnight Wish S-Boxes

Blue Midnight Wish hash function has the following bijective components (S-boxes) that are the subject of interest in this paper:

$$BMW224/256: \begin{cases} s_0(x) = SHR^1(x) \oplus SHL^3(x) \oplus ROTL^4(x) \ \ \oplus ROTL^{19}(x) \\ s_1(x) = SHR^1(x) \oplus SHL^2(x) \oplus ROTL^8(x) \ \ \oplus ROTL^{23}(x) \\ s_2(x) = SHR^2(x) \oplus SHL^1(x) \oplus ROTL^{12}(x) \oplus ROTL^{25}(x) \\ s_3(x) = SHR^2(x) \oplus SHL^2(x) \oplus ROTL^{15}(x) \oplus ROTL^{29}(x) \end{cases} \tag{8}$$

$$BMW384/512: \begin{cases} s_0(x) = SHR^1(x) \oplus SHL^3(x) \oplus ROTL^4(x) \ \ \oplus ROTL^{37}(x) \\ s_1(x) = SHR^1(x) \oplus SHL^2(x) \oplus ROTL^{13}(x) \oplus ROTL^{43}(x) \\ s_2(x) = SHR^2(x) \oplus SHL^1(x) \oplus ROTL^{19}(x) \oplus ROTL^{53}(x) \\ s_3(x) = SHR^2(x) \oplus SHL^2(x) \oplus ROTL^{28}(x) \oplus ROTL^{59}(x) \end{cases} \tag{9}$$

where $ROTL^n(x)$ means rotation of the variable x to the left for n positions and $SHL^n(x)$ means shifting of the variable x to the left for n positions (variables are 32-bit long for BMW224/256 and they are 64-bit long for BMW384/512).

By simple application of the Definition 1 we can obtain the following result:

Corollary 6.

$$BMW224/256: \begin{cases} C\left(\mathbf{s_0}^{-1}\right) = 524 \\ C\left(\mathbf{s_1}^{-1}\right) = 528 \\ C\left(\mathbf{s_2}^{-1}\right) = 528 \\ C\left(\mathbf{s_3}^{-1}\right) = 528 \end{cases} \tag{10}$$

$$BMW384/512: \begin{cases} C\left(\mathbf{s_0}^{-1}\right) = 2080 \\ C\left(\mathbf{s_1}^{-1}\right) = 2080 \\ C\left(\mathbf{s_2}^{-1}\right) = 2080 \\ C\left(\mathbf{s_3}^{-1}\right) = 2080 \end{cases} \tag{11}$$

□

Although S-boxes in Blue Midnight Wish have four operations (compared to the three of SHA-2), it comes as a little surprise that the maximal value of $C\left(\mathbf{s}^{-1}\right)$ for $n = 32$ and $n = 64$ for the types of S-boxes defined in Blue Midnight Wish is not much bigger than the corresponding maximal values for SHA-2. That is easily checkable fact by a simple exhaustive search of all possible S-boxes of the type defined in Blue Midnight Wish.

Corollary 7. *If the function* $\mathbf{s} : \{0,1\}^{32} \rightarrow \{0,1\}^{32}$ *is defined as*

$$s(x) = SHR^{r_1}(x) \oplus SHL^{r_2}(x) \oplus ROTL^{r_3}(x) \oplus ROTL^{r_4}(x) \equiv \mathbf{s} \cdot x \tag{12}$$

where $0 \leq r_1 < 32$*,* $0 \leq r_2 < 32$*,* $0 \leq r_3 < r_4 < 32$*, then* $Max(C(\mathbf{s}^{-1})) = 528$*.*

□

Corollary 8. *If the function* $\mathbf{s} : \{0,1\}^{64} \rightarrow \{0,1\}^{64}$ *is defined as*

$$s(x) = SHR^{r_1}(x) \oplus SHL^{r_2}(x) \oplus ROTL^{r_3}(x) \oplus ROTL^{r_4}(x) \equiv \mathbf{s} \cdot x \tag{13}$$

where $0 \leq r_1 < 64$*,* $0 \leq r_2 < 64$*,* $0 \leq r_3 < r_4 < 64$*, then* $Max(C(\mathbf{s}^{-1})) = 2080$*.*

□

Corollaries 6, 7 and 8 show that we chose 7 S-boxes with maximal $C(\mathbf{s}^{-1})$ value and one with a value which is very near to the maximum. More precisely, our design criteria were the following:

- Logical functions s_i, $i = 0, \ldots, 3$, are bijections in $\{0,1\}^{32} \rightarrow \{0,1\}^{32}$ (resp. in $\{0,1\}^{64} \rightarrow \{0,1\}^{64}$) i.e. they are S-boxes.
- They have different pairs of 1-bit, 2-bits or 3-bits shifts to the left and to the right.
- They have different pairs of rotations to the left, in such a way that one rotation is less than $w/2, w = 32, 64$, and the other rotation is bigger than $w/2$.
- The values of the rotations that are less than $w/2$ are in the interval of ± 2 (resp. ± 4) around numbers $\{2, 6, 10, 14\}$ (resp. $\{4, 12, 20, 28\}$).

- The values of the rotations that are bigger than $w/2$ are in the interval of ± 2 (resp. ± 4) around numbers $\{18, 22, 26, 30\}$ (resp. $\{36, 42, 50, 58\}$).
- The values $C\big(\mathbf{s}_i{}^{-1}\big),\ i = 0, \ldots, 3$, to be the maximal possible (or very close to the maximal value).

By computer search we have found hundreds of such bijections and from them we have chosen the eight particular functions s_0, s_1, s_2 and s_3 (four for BMW224/256 and four for BMW384/512).

4 Conclusions and Future Work

The design principles of SHA-2 family of hash functions are still kept as a classified NSA information. In the open literature there have been several attempts to analyze those design principles.

In the design of Blue Midnight Wish cryptographic hash function as a SHA-3 candidate, several bijective components (S-boxes) have been chosen with properties that are similar to the properties of S-boxes in SHA-2.

The observations presented in this paper probably open more new questions than close some. One obvious thing that have to be done in the next period would be to establish firm mathematical connection between our defined measure $Max(C(\mathbf{s}^{-1}))$ and the theory of computational asymmetry.

References

1. Rivest, R.: The MD4 message-digest algorithm, Request for Comments (RFC) 1320, Internet Activities Board, Internet Privacy Task Force (April 1992)
2. Rivest, R.: The MD5 message-digest algorithm, Request for Comments (RFC) 1321, Internet Activities Board, Internet Privacy Task Force (April 1992)
3. FIPS 180-1, Secure Hash Standard, Federal Information Processing Standards Publication 180-1, U.S. Department of Commerce/NIST, National Technical Information Service, Springfield, Virginia (April 1995)
4. FIPS 180-2, Secure Hash Standard, Federal Information Processing Standards Publication 180-2, U.S. Department of Commerce/NIST, National Technical Information Service, Springfield, Virginia (August 2002)
5. den Boer, B., Bosselaers, A.: An attack on the last two rounds of MD4. In: Feigenbaum, J. (ed.) CRYPTO 1991. LNCS, vol. 576, pp. 194–203. Springer, Heidelberg (1992)
6. den Boer, B., Bosselaers, A.: Collisions for the compression function of MD-5. In: Helleseth, T. (ed.) EUROCRYPT 1993. LNCS, vol. 765, pp. 293–304. Springer, Heidelberg (1994)
7. Vaudenay, S.: On the need for multipermutations: Cryptanalysis of MD4 and SAFER. In: Preneel, B. (ed.) FSE 1994. LNCS, vol. 1008, pp. 286–297. Springer, Heidelberg (1995)
8. Dobbertin, H.: Cryptanalysis of MD4. J. Cryptology 11, 253–271 (1998)
9. Chabaud, F., Joux, A.: Differential collisions in SHA-0. In: Krawczyk, H. (ed.) CRYPTO 1998. LNCS, vol. 1462, pp. 56–71. Springer, Heidelberg (1998)
10. Biham, E., Chen, R.: Near-collisions of SHA-0. Cryptology ePrint Archive, Report 2004/146 (2004), http://eprint.iacr.org/2004/146

11. Wang, X., Lai, X., Feng, D., Chen, H., Yu, X.: Cryptanalysis of the Hash Functions MD4 and RIPEMD. In: Cramer, R. (ed.) EUROCRYPT 2005. LNCS, vol. 3494, pp. 1–18. Springer, Heidelberg (2005)
12. Wang, X., Yu, H.: How to Break MD5 and Other Hash Functions. In: Cramer, R. (ed.) EUROCRYPT 2005. LNCS, vol. 3494, pp. 19–35. Springer, Heidelberg (2005)
13. Wang, X., Yu, H., Yin, Y.L.: Efficient Collision Search Attacks on SHA-0. In: Shoup, V. (ed.) CRYPTO 2005. LNCS, vol. 3621, pp. 1–16. Springer, Heidelberg (2005)
14. Wang, X., Yin, Y.L., Yu, H.: Finding Collisions in the Full SHA-1. In: Shoup, V. (ed.) CRYPTO 2005. LNCS, vol. 3621, pp. 17–36. Springer, Heidelberg (2005)
15. Gilbert, H., Handschuh, H.: Security analysis of SHA-256 and sisters. In: Matsui, M., Zuccherato, R.J. (eds.) SAC 2003. LNCS, vol. 3006, pp. 175–193. Springer, Heidelberg (2004)
16. Hawkes, P., Paddon, M., Rose, G.G.: On corrective patterns for the SHA-2 family, Cryptology ePrint Archive, Report 2004/207 (2004)
17. Yoshida, H., Biryukov, A.: Analysis of a SHA-256 variant. In: Preneel, B., Tavares, S. (eds.) SAC 2005. LNCS, vol. 3897, pp. 245–260. Springer, Heidelberg (2006)
18. Matusiewicz, K., Pieprzyk, J., Pramstaller, N., Rechberger, C., Rijmen, V.: Analysis of simplified variants of SHA-256. In: Proceedings of WEWoRC 2005, vol. LNI P-74, pp. 123–134 (2005)
19. Mendel, F., Pramstaller, N., Rechberger, C., Rijmen, V.: Analysis of step-reduced SHA-256. In: Robshaw, M.J.B. (ed.) FSE 2006. LNCS, vol. 4047, pp. 126–143. Springer, Heidelberg (2006)
20. Hölbl, M., Rechberger, C., Welzer, T.: Finding message pairs conforming to simple SHA-256 characteristics: Work in Progress. In: Western European Workshop on Research in Cryptology - WEWoRC 2007, Bochum, July 4-6, pp. 21–25 (2007), http://www.hgi.rub.de/weworc07/PreliminaryConferenceRecord.pdf
21. Nikolìc, I., Biryukov, A.: Collisions for Step-Reduced SHA-256. In: Nyberg, K. (ed.) FSE 2008. LNCS, vol. 5086, pp. 1–15. Springer, Heidelberg (2008)
22. Indesteege, S., Mendel, F., Preneel, B., Rechberger, C.: Collisions and other Non-Random Properties for Step-Reduced SHA-256. Cryptology ePrint Archive, Report 2008/131 (2008), http://eprint.iacr.org/
23. NIST, First Cryptographic Hash Workshop, October 31 - November 1 (2005), Second Cryptographic Hash Workshop, August 24-25 (2006), http://csrc.nist.gov/groups/ST/hash/first_workshop.html, http://csrc.nist.gov/groups/ST/hash/second_workshop.html
24. NIST Tentative Timeline for the Development of New Hash Functions, http://csrc.nist.gov/groups/ST/hash/timeline.html
25. NIST, SHA-3 First Round Candidates, http://csrc.nist.gov/groups/ST/hash/sha-3/Round1/submissions_rnd1.html
26. NIST, SHA-3 Second Round Candidates, http://csrc.nist.gov/groups/ST/hash/sha-3/Round2/submissions_rnd2.html
27. Gligoroski, D., Klima, V., Knapskog, S.J., El-Hadedy, M., Amundsen, J., Mjølsnes, S.F.: Cryptographic Hash Function BLUE MIDNIGHT WISH. Submission to NIST (2008)
28. Birget, J.-C.: One-way permutations, computational asymmetry and distortion”, arXiv:0704.1569v1 [math.GR] (2007), http://arxiv.org/abs/0704.1569v1
29. Klima, V.: Tunnels in Hash Functions: MD5 Collisions Within a Minute, Cryptology ePrint Archive, Report 2006/105 (2006), http://eprint.iacr.org/

Optimization of Adaptive Petri-Net Grid Genetic Algorithm Workflows

Boro Jakimovski, Dragan Sahpaski, and Goran Velinov

Institute of Informatics, Faculty of Sciences and Mathematics
Ss. Cyril and Methodius University, Skopje, Macedonia
{boroj,dragans,goranv}@ii.edu.mk

Abstract. In this paper we present an analysis methodology for the possible improvements of the performance of Adaptive Petri-Net Grid Genetic Algorithm workflow. Genetic Algorithms are very powerful optimization technique that is easily parallelized using different approaches which makes it ideal for the Grid. The High Level Petri-Net workflow model greatly outperforms currently available DAG workflow model available in gLite Grid middleware. Using the flexibility of the High Level Petri-Net workflows we have designed an adaptive workflow that overcomes the heterogeneity and unpredictability of the Grid infrastructure, giving users better and more stable execution times than formerly used DAG workflows. The performance of the Petri-Net Grid Genetic Algorithm is analyzed using several parameters that change the behavior of the optimization. The performance is measured as a shortening of the overall execution time of the workflow in the process of searching for a solution with suitable quality. In the course of the analysis we defined a stable measurement of the quality of the solution used in the experiments. The experimental results obtained by Genetic Algorithm optimization of performance of the Data Warehouse design have shown the unexpected interesting influence some parameters have over the optimization time for obtaining the same quality level.

Keywords: Genetic Algorithms, High level Petri-nets, Grid Workflows.

1 Introduction

Workflows programming model represents very appealing approach in Grid scientific computations, because it allows representation of large scientific computational processes as a high level graphs. Different approaches have emerged for representation of Grid workflows, having different level of complexity and capabilities.

Currently most widely used workflow model on the gLite Grid middleware is DAG (Direct Acyclic Graph). It enables easy representation of simple scientific workflows, allowing only acyclic data dependencies in the graph of processes and thus enabling sequence and parallelism constructs. Even though DAG workflows in theory should enable choice, gLite DAG workflows do not allow for a choice construct. This limits the ability of scientists to design more complex workflows containing choices and iterations. Different Grid projects allow the execution of more complex workflows relying on different workflow models. It has been shown that the Petri-Net formalism for workflow description outperforms other modeling formalisms, mainly due to its formal state-based structure and availability of analysis techniques [1]. GWorkflowDL [7] is an XML style language for representation of High Level Petri-Nets (HLPN) for scientific Grid workflows. It has not been adopted by gLite, but simple Workflow Management Service (WfMS) for execution of GWorkflowDL workflows on gLite is in development.

Evolutionary Algorithms (EA) are a computational model inspired by the natural process of evolution. They have been successfully used for solving complex optimization problems. Genetic Algorithms (GA), a subclass of EA, search for potential solution by encoding the data into a chromosome-like structure. The search is done over a set of chromosomes (population) with repetitive application of recombination, mutation and selection operators until a certain condition is reached. One repetition is called a generation.

Usually the search for a solution using Genetic Algorithm is a long and computationally intensive process. Researchers running hundreds of optimizations, before they reach desired results are in great need for CPU power. Fortunately the Genetic Algorithms are easily parallelized using data partitioning of the population among different processes. This kind of parallelism ensures close to linear speedup, and sometimes super-linear speedup. Over the past years many variants of parallelization techniques have been exploited for parallelization of Genetic Algorithms [3][12].

Grid genetic algorithms have been efficiently used for solving different problems [6] [8]. Our previous research in the field of Grid Genetic Algorithms was focused on the development, analysis and optimization of different Grid workflow modeling techniques, starting from gLite DAG workflows [9][15], followed by the Adaptive Grid Genetic Algorithms based on the High Level Petri-Net GWorkflowDL model that copes with the unpredictability and heterogeneity of the Grid infrastructure and thus obtains much more stable and shorter execution time than previously used DAG workflows. [10]

The evaluation of the performance of the adaptive Grid workflow algorithm was done by optimization of Data Warehouse performance using genetic algorithms. The goal of the optimization process is to select the optimal state of the implementation schema (optimal Data Warehouse design) in order to minimize the workload processing cost. Our previous research in the field of Data Warehouse performance optimization by Grid workflows has shown great gain in performance compared to standard Genetic Algorithms. The analysis of the results has shown that the speedup factor is greatly influenced by the level of concurrency introduced in the workflow for a given problem size, due to instability and unpredictability of the Grid [15].

In this paper we present a more in-depth analysis of the adaptive workflow grid genetic algorithm. The results of the analysis show the possibility of optimization of the total execution time by fine tuning workflow parameters.

The rest of this paper is organized as follows. In Section 2, we present the Grid genetic algorithms and the Petri-Net workflow model. In Section 3, we focus on the analysis and optimization of the Adaptive Grid Genetic Algorithm based on the Petri-Net workflow model. The experimental results obtained by optimization of Data

Warehouse problem for the performance measurements of the proposed workflow tuning parameters of the Adaptive Grid Genetic Algorithm approach is presented in Section 4. Finally, Section 5 concludes this paper and gives future development issues.

2 Optimization of Adaptive Grid Genetic Algorithms

The Parallel Genetic Algorithms (PGAs) are extensions of the single population GA. The well-known advantage of PGAs is their ability to perform speciation, a process by which different subpopulations evolve in diverse directions simultaneously. They have been shown to speed up the search process and to attain higher quality solutions on complex design problems [4][14].

There are three major classes of Parallel Genetic Algorithms: Master-slave, Cellular and Island. Master-slave parallelization uses single population of chromosomes and parallelizes only the chromosome evaluation part of the optimization. This makes it suitable for usage in parallel environments that have shared memory. Cellular Parallel Genetic Algorithms also consist of single chromosome population, but the computation can be partially structured. This is mostly suitable for massively parallel systems, consisting of large number of processing elements organized in a topology, which is followed by the Parallel Genetic Algorithm. Most widely used and most sophisticated Parallel Genetic Algorithms is the Island PGA, or in other words Multi-population PGA. This approach enables parallel nearly-independent execution of populations. The only connection between the populations is occasional migration of chromosomes. This kind of Parallel Genetic Algorithm is suitable for message passing parallelism environments.

The nature of the Grid makes it best suited for Island Parallel Genetic Algorithm for achieving high performance parallelism. We chose to investigate the implementation of Grid genetic algorithms using Grid workflows for work distribution. Grid workflows represent a network of interconnected Grid jobs where interconnected jobs are data dependent, i.e. data output from one job is feeded into the dependent jobs for further processing. This model of parallel execution allows easy, efficient and more natural use of the Grid for data parallelization problems that can be easily divided into parallel independent jobs. When we gridify the Genetic Algorithms, each job in the workflow represents single population, i.e. island, which is iterated several generations. Iteration of one generation includes breeding, mutation and selection on the population. We name such jobs Breeder.

The definition of the workflow that will implement the distributed execution of the Grid genetic algorithm can be defined in many different ways. When using DAG workflows we defined the workflow in epochs, where we wait for all started islands to finish the breeding process for certain number of iterations, i.e. end of an Epoch. Then Grid Genetic Algorithm needs to migrate population between the islands. This was implemented in the workflow as a separate workflow job, Migrator that joins output populations from several jobs, and after migration of genetic material, sends the mixed populations to a new set of Breeder jobs, i.e. start of a new Epoch. The whole optimization Grid workflow is defined as a several epochs of breeding, where at the end we collect the final populations as a final result, Collector job [9]. We addressed the disadvantages of the static DAG nature by using adaptive Genetic Algorithm using Petri-net workflows [10]. First optimization step of the static DAG Grid Genetic Algorithm was by joining the Migrator jobs with their following workflow dependent Breeder jobs. This process eliminates the scheduling of large number of very short jobs and thus shortens the time greatly. Even with this optimization we saw that the achieved level of concurrency and speedup was far from the desired level. After the analysis of the workflow jobs execution times we concluded that a stalling job (long queue time) were the main reason for the degraded the level of concurrency of the workflow, leading to an large and instable execution time with a big deviation from the mean estimated execution time. This unpredictability was addressed by using more powerful workflow expression formalism – High Level Petri-net workflows.

The High Level Petri-Net (HLPN) model extends classical Petri-Net model with features that make them more suitable for workflow representation [1]. High Level Petri-Nets allow for nondeterministic and deterministic choice, simply by connecting several transitions to the same input place and annotating edges with conditions. For incoming edges, variable names are used as edge expressions, which assign the token value obtained through this edge to a variable of that name. Additionally, each transition can have a set of boolean condition functions. A transition can only fire if all of its conditions evaluate to true for the input tokens. Another advantage over DAG is that DAGs only have a single node type, which means that data flowing through the net cannot be modeled easily. In contrast, Petri Nets are able to model the state of the program execution by tokens flowing through the net.

Grid Petri-net workflows represent standard Petri-net workflows where the transitions can represent Grid Job submission and tokens represent data flowing from job to job. This makes the Petri-net Grid workflow model very appealing to scientists as it gives better mapping of real scientific processes to Grid workflows.

Many efforts have been done in order to propose a Grid workflow description language based on the Petri Nets formalism. GWorkflowDL is a language developed by the Fraunhofer FIRST research group an XML-based language which uses the High Level Petri Nets (HLPN) modeling formalism allowing extensions with user defined operation types [2][7]. In the course of EGEE II project a development of WfMS was started as a Workflow engine to support GWorkflowDL workflows on gLite middleware [13]. In our research we used the WfMS for gLite for the execution of the Adaptive Grid Genetic Algorithm workflow. In order to define the Adaptive Grid Genetic Algorithm using Petri-net workflow model we used the ability to define a non-deterministic connection between jobs. This means that after the submission of the initial Epoch of Breeder jobs, the workflow does not need to specify which specific jobs need to finish, for the migration to take place. Hence the new submission of Migration/Breeder jobs will not be influenced by the stalling jobs which was the case in DAG workflows. The Petri-net graph of the Adaptive Grid Genetic algorithm is shown in Fig. 1.

Another ability of the Petri-net workflow model is the availability of conditional loops, which can enable optimizations to be executed until certain threshold is met. This allows for genetic optimization using non fixed number of islands, which enables search for the solution with a certain level of quality.

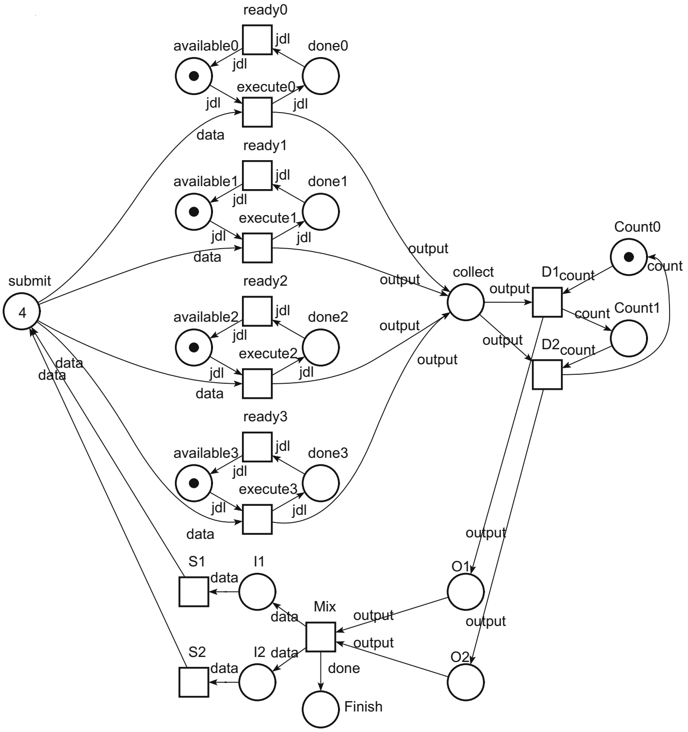


Fig. 1. Adaptive Grid Genetic Algorithm Petri-net workflow for 4 concurrent islands and MF=2

3 Analysis of the Adaptive Grid Genetic Algorithm

Prior to the analysis of the different parameter influence of the overall execution time we need to define a solution quality assessment methodology, i.e. definition of a measuring technique that will determine that the solution has converged and we do not need to further search for a better one. This poses a great challenge for several reasons. First the nature of the adaptive algorithm is that it needs to assess the solutions very frequently and with partial information, i.e. when several jobs finish the workflow assesses the solution and decides if it will continue with new job submissions or it will wait for the others to finish. Secondly, if we have all population results it is very hard to assess the convergence of the solution.

In order to select the proper solution quality measurement we executed many workflows with a large number of jobs in order to see how is the process of convergence towards the solution with respect to the finishing times of the Grid jobs. We noticed that the solution quality tends to increase in steps shorter or in the order of the level of concurrency in the workflow. This means that when analyzing the level of quality in the solution one needs to analyze only the several last finished populations *k*. The amount of the population analyzed should be at least the level of concurrency in the workflow, but not the whole set of finished populations. We investigated several techniques in order to assess if the increase of the solution quality is significant enough for the search to continue. We chose to use the following measurement:

$$m = \frac{S_x}{\overline{X}} \tag{3.1}$$

where X is the series obtained in the following manner. After each batch of populations/jobs finish (migration factor), we record the best solution until each population ordered by time, and insert the best solution as a new item in the series. If the series is longer than the predefined value of *k*, we remove the oldest items in the sample in order for X to have always maximum *k* elements. The choice of this measurement is to evaluate the deviation of the result not to vary too much in respect to its value. In our further research of the parameters we will choose m to be lower then 0.1 (obtained for our experiments by experimental evaluation).

Adaptive Grid Genetic Algorithm has several parameters that can be used for changing the behavior of the workflow and thus obtains different execution time and different quality of the solution. These parameters are:

- Migration Factor – number of populations that exchange genetic material once they finish an epoch. This parameter will influence two aspects of the optimization. First is the overall execution time. In our previous work we have concluded that the overall workflow execution time greatly increases as the factor increases due to the bigger amount of stalling jobs influence on the submission. On the other hand bigger migration factor greatly influences the convergence of towards the solution. This is why we consider investigating this correlation.
- Number of Concurrent Islands – number of populations that are submitted to the grid in parallel. Number of concurrent islands potentially enables bigger concurrency in the workflow execution. Nevertheless, the availability of resources and splitting the genetic algorithm in too many populations might have decreasing factor on the performance.
- Migration Strategy – selection of different migration strategy might speed up solution search and achieve better quality. We choose the strategy where a certain number of chromosomes of each population is migrated to all other populations that enter the migration process. This parameter is called the Migration Step. By default we choose to use the elitism strategy where we migrate only the best chromosome of the populations that enter the migration process. We decided not to experiment with this parameter according to experimental results shown by [5].
- Number of Iterations per Population – this is the number of generations each population executes in an epoch, i.e. in a single Grid job. This parameter influences the performance from two aspects. Larger number of generations increases the execution time per process/job, making the Grid overheads look insignificantly small. On the other hand making too many generations without migrations might produce worse results.

4 Experimental Results

We evaluated the influence of the different parameters over the overall optimization time of the Adaptive Grid Genetic Algorithm by optimizing Data Warehouse performance. The influence was measured only in the course of time optimization, i.e. shortening the overall execution time of the workflow while obtaining a solution that satisfies the quality measurement level. The mentioned optimization problem was described in great details in our previous work [15][10]. As in our previous work, the experiments were implemented in Java using JGAP [11] library and was executed on SEE-GRID infrastructure.

As defined in the previous section the experiments were evaluated using three out of four variable parameters. The values used for evaluation of the parameters were: 3-5-7 for the Migration Factor (MF), 10-20-40 for the Number of concurrent islands (N) and 50-100-400 for the number of iterations per population (I). Usual execution time of one population of 100 iterations is around 20 minutes.

In Fig. 2 - Fig. 4 we can see three graphs representing the obtained results from the parameter evaluation. They represent the experimental optimization time for obtaining the solution for 50, 100 and 400 iterations per population/job respectively.

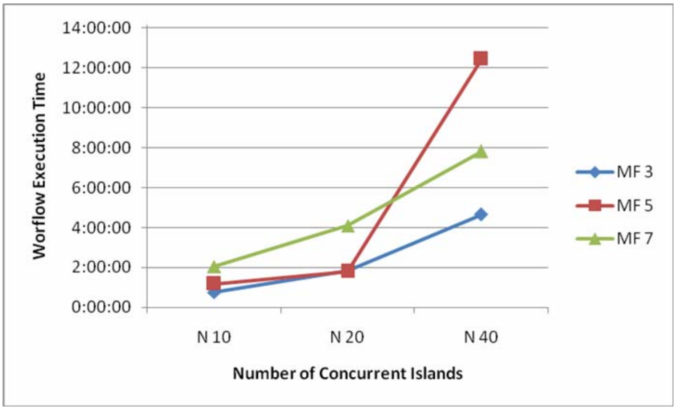


Fig. 2. Evaluation of the parameter N and MF influence over optimization time for I=50

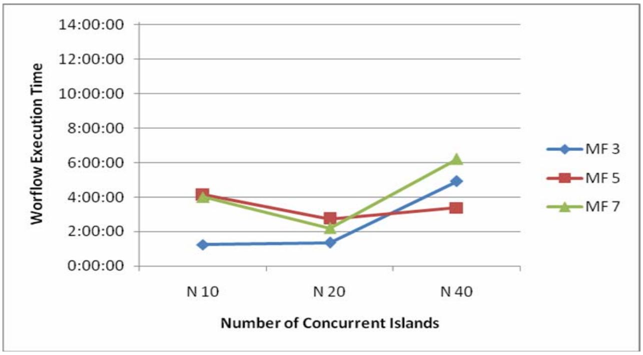


Fig. 3. Evaluation of the parameter N and MF influence over optimization time for I=100

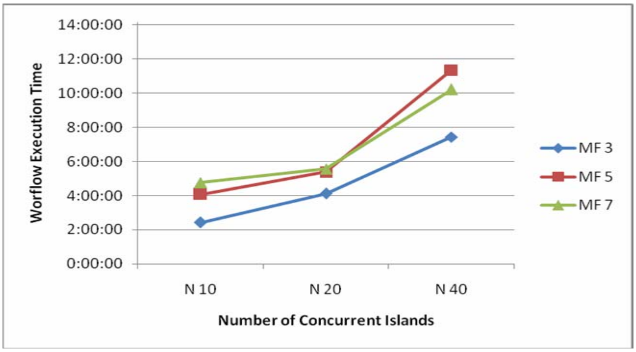


Fig. 4. Evaluation of the parameter N and MF influence over optimization time for I=400

The three different lines represent different migration factors mentioned above, as for the X-axis represents the concurrency level, i.e. number of parallel islands. As it can be seen from the graphs the best solution is always obtained using the lowest parameters, i.e. 10 parallel jobs, using migration factor 3, number of iterations 50. The best optimization time was not very much influenced by number of iterations per job. This shows that even though bigger concurrency level and bigger number of iterations per population should increase convergence speed, it is greatly influenced badly from the long queue times of too many jobs.

As mentioned in previous chapter we use competition measurement of 0.05. The reason for this can be seen from Fig. 5 that shows the effect on quality of the solution as opposed to different values of *m*. The values in the graph shows mean percentage of difference from the best achieved solution, together with standard deviation from the mean percentage. The best achieved solution was obtained as we discarded the quality measurement technique for stopping the optimization, and let the optimization work for 200 populations. It can be seen that this method ensures good evaluation of the quality of the result even for very low values of *m*.

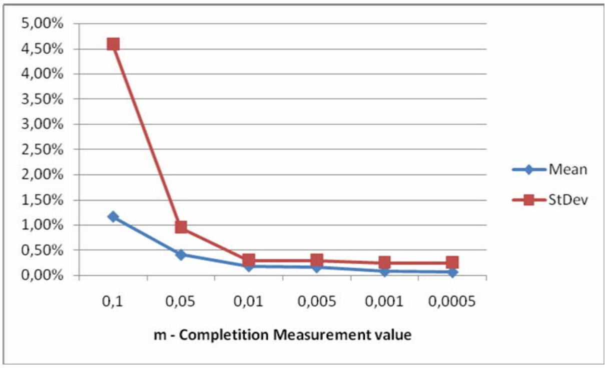


Fig. 5. Evaluation of the influence of Competition Measurement *m* over the obtained solution

5 Conclusion

In this paper we have analyzed the possibilities of optimization of the previously defined and researched Adaptive Grid Genetic Algorithm. The defined performance parameters were experimentally analyzed for their effect over the overall execution time of the Adaptive Grid Genetic Algorithms in the search for a good quality solution. As a consequence we have defined a good measurement for evaluation of the quality of the solution that greatly shortens the overall execution time. The experimental results show that despite the expected increase in speedup of solution convergence using bigger migration factors, higher number of iterations and using more concurrency, the influence of the Grid instability factors is too big and that the best solution is obtained using low migration, low concurrency level and low number of iterations per population. This is not a general case, since other factors that are relative to the optimization also influence the performance. Some of them are Grid related (available Grid resources), some are problem related (depending on the nature of the problem). Nevertheless the methodology covered in this paper enables researchers using Adaptive Grid Genetic Algorithm to evaluate their specific case of optimization and determine the parameters that give best performance.

Future research needs to be done concerning other aspects of optimization of performance. In this paper we only addressed the workflow and Genetic Algorithm aspects that influence the performance. The best performance optimization can be archived by addressing the main reason for performance degradation – long Grid job queuing times.

Acknowledgment

The work in this paper is partially supported by the FP7 SEE-GRID-SCI and NATO ICS.EAP.CLG.983334 projects.

References

1. Aalst, W.: The Application of Petri Nets to Workflow Management. The Journal of Circuits, Systems and Computers 8(1), 21–66 (1998)
2. Alt, M., et al.: Using High Level Petri-Nets for Describing and Analysing Hierarchical Grid Workflows. In: Proc. of the CoreGRID Integration Workshop, Pisa, Italy (2005)
3. Cantu-Paz, E.: A Survey of Parallel Genetic Algorithms. Calculateurs Paralleles, Reseaux et Systems Repartis 10(2), 141–171 (1998)
4. Cui, J., Fogarty, T.C., Gammack, J.G.: Searching Databases Using Parallel Genetic Algorithms on a Transputer Computing Surface. Future Generation Computer Systems 9(1), 33–40 (1993)
5. Golub, M.: Improving the Efficiency of Parallel Genetic Algorithms. Ph.D. thesis, Zagreb University, Croatia (2001)
6. Herrera, J., Huedo, E., Montero, R.S., Llorente, I.M.: A Grid-Oriented Genetic Algorithm. In: Sloot, P.M.A., Hoekstra, A.G., Priol, T., Reinefeld, A., Bubak, M. (eds.) EGC 2005. LNCS, vol. 3470, pp. 315–322. Springer, Heidelberg (2005)

7. Hoheisel, A., Der, U.: An XML-based Framework for Loosely Coupled Applications on Grid Environments. In: Sloot, P.M.A., Abramson, D., Bogdanov, A.V., Gorbachev, Y.E., Dongarra, J., Zomaya, A.Y. (eds.) ICCS 2003. LNCS, vol. 2657, pp. 245–254. Springer, Heidelberg (2003)
8. Imade, H., Morishita, R., Ono, I., Ono, N., Okamoto, M.: A Grid-Oriented Genetic Algorithm Framework for Bioinformatics. New Generation Computing 22(2), 177–186 (2004)
9. Jakimovski, B., Cerepnalkoski, D., Velinov, G.: Framework for Workflow Gridication of Genetic Algorithms in Java. In: Bubak, M., van Albada, G.D., Dongarra, J., Sloot, P.M.A. (eds.) ICCS 2008, Part III. LNCS, vol. 5103, pp. 463–470. Springer, Heidelberg (2008)
10. Jakimovski, B., Sahpaski, D., Velinov, G.: Performance improvement of Genetic Algorithms by Adaptive Grid Workflows. In: Proc. of the 11th Intl Symposium on Symbolic and Numeric Algorithms for Scientific Computing. IEEE Press, Timisoara (2009)
11. Meffert, K.: JGAP - Java Genetic Algorithms and Genetic Programming Package (2009), http://jgap.sf.net
12. Nowostawski, M., Poli, R.: Parallel Genetic Algorithm Taxonomy. In: Proc. of the Third International Conference on Knowledge-Based Intelligent Information Engineering Systems (KES 1899), Adelaide, pp. 88–92 (1899)
13. Pellegrini, S., Giacomini, F.: Design of a Petri Net-Based Workflow Engine. In: Proc. of the 3rd Intl. Conference on Grid and Pervasive Computing Workshops 2008, pp. 81–86 (2008)
14. Sena, G.A., Megherbi, D., Isern, G.: Implementation of a Parallel Genetic Algorithm on a Cluster of Workstations: Travelling Salesman Problem, a Case Study. Future Generation Computer Systems 17(4), 477–488 (2001)
15. Velinov, G., Jakimovski, B., Cerepnalkoski, D., Kon-Popovska, M.: Framework for Improvement of Data Warehouse Optimization Process by Workflow Gridification. In: Proc. of the 12th Conference on Advances in Databases and Information Systems, ADBIS 2008. Pori, Finland, pp. 295–304 (2008)

Massively Parallel Seismic Data Wavelet Processing Using Advanced Grid Workflows

Ljupco Jordanovski, Boro Jakimovski, and Anastas Misev

University Sts Cyril and Methodius, Faculty of Natural Sciences & Mathematics Skopje, Macedonia
ljordanovski@gmail.com, boroj@ii.edu.mk, anastas@ii.edu.mk

Abstract. Leveraging the huge seismic data collections can be a quite challenging process, especially if the available data comes from large number of sources. Computing Grids enable such processing, giving the users necessary tools to share the data from various countries and sources. Processing this data not only gives results related to the earthquakes themselves, but also it reveals the geological features of the observed regions. Using the gLite base Grid, we propose a framework for massively parallel wavelet data processing of the seismic waveforms using advanced Grid workflows. Such workflows enable users to use the power of the Grid more easily and to achieve better performance. In the process of the data processing we use seamlessly several different grid services (AMGA, LFC ...) to locate the necessary data and to extract the needed information. The Grid application uses waveform data from several earthquakes from the same recording station. For the processing we use continuous wavelet transformation in order to capture the characteristics of the earth crust following the path from the earthquake origin towards the station. These features are recorded and later are classified using pattern matching to identify important characteristics of some specific seismic region as seen from that specific station.

1 Introduction

Wavelet theory has matured in past years as a new mathematical tool for time series analysis [3] [5] [7]. However, although it seems quite natural to be applied in seismology, it seems that it is in its initial stage [2].

Having in mind that the input data for such processing is based on the waveform data recorded by the seismic sensors, which represents a huge volume of data, it is necessary to facilitate some mechanisms and methodologies to cope with its storage and processing. Computational Grids present a powerful and most suited processing tool to address both the distributed storage and the high level of processing needed.

Since the processing of the data is not trivial, we propose the usage of complex workflows based on the Petri-net Grid Workflows. The High Level Petri-Net workflow model greatly outperforms currently available DAG workflow model available in gLite Grid middleware. DAG workflows only allow acyclic data dependencies in the graph of processes that enables sequence and parallelism construct. This limitation allows only representation of simple scientific workflows, which limits the ability of scientists to design more complex workflows containing choices and iterations [1].

The High Level Petri-Net (HLPN) model extends classical Petri-Net model with features that make them more suitable for workflow representation [1]. High Level Petri-Nets allow for nondeterministic and deterministic choice, simply by connecting several transitions to the same input place and annotating edges with conditions. For incoming edges, variable names are used as edge expressions, which assign the token value obtained through this edge to a variable of that name. Additionally, each transition can have a set of Boolean condition functions. A transition can only fire if all of its conditions evaluate to true for the input tokens. Another advantage over DAG is that DAGs only have a single node type, which means that data flowing through the net cannot be modeled easily. In contrast, Petri Nets are able to model the state of the program execution by tokens flowing through the net.

Grid Petri-net workflows represent standard Petri-net workflows where the transitions can represent Grid Job submission and tokens represent data flowing from job to job. This makes the Petri-net Grid workflow model very appealing to scientists as it gives better mapping of real scientific processes to Grid workflows.

GWorkflowDL [11] is an XML style language for representation of High Level Petri-Nets (HLPN) for scientific Grid workflows. It has not been adopted by gLite, but simple Workflow Management Service (WfMS) for execution of GWorkflowDL workflows on gLite is in development.

2 Wavelet Analysis of the Seismic Data

Analysis of the earthquake data can reveal different characteristics regarding the earthquakes origin and the waves travelling through earths crust. After exploring several signal analysis techniques we noticed that making a continuous or discrete wavelet transformations and plotting the results in coordinate system, scales versus time, one could notice striking similarity of the wavelet images, coming from the same source region between different seismic records. On the other hand the images had noticeable difference for records of earthquakes occurred in different source region. It is assumed that those similar image patterns are due to same underlying geological setting while the differences (usually for smaller scale) are due to different source mechanism and finer geological structures. In the first approximation of geological structure, similarities of the image patterns in domain of large scale are noticeable even for the records from different source regions.

With massive processing of earthquake records one could define: Common features of the propagation path for the given seismic source region or to define empirical transfer function of the media (Green function); Calculation of the artificial seismograms; Determine the source region based on a single earthquakes record; Determine the more realistic attenuation curve of the selected feature (parameter), very much needed in seismic hazard and risk analysis; Mapping (coding) of the given earthquake prone region in terms of selected parameters; Seismic source parameters.

Apart from above stated goals other possible benefits could be:

- Better noise removal from the seismic records.
- Possible better and more accurate different phase identification.

In order to facilitate this research a great computational power is needed in order to extract as many features as possible from as many as possible seismic regions. As the sample of processed earthquakes gets bigger and bigger, a better trained classificators will be developed. A solution for the great computational and storage needs of this research lies in Computational Grids. As in our previous work, we use gLite Grid middleware and implement the application on the SEE-GRID infrastructure.

2.1 Wavelet Processing

During developing the main core of the algorithm regarding wavelet data processing we have realized that certain parameters of the records should be the same for every processing:

- Time span before the first wave phase (usually P-phase) reached recording station.
  - It could be constant or
  - It could be time of the earthquake occurrence in hypocenter. (For the regional earthquakes for distance less than 300 km it is less than 30 to 40 sec.)
- Total duration of the record should be equal, allowing small portion of data after seismic waves passed the station and records reached noise level.
- Equal time sampling interval (delta t).
- It is also desirable but not necessarily if all records are filtered with same filter so that frequency band is equal for all records.

We have focused on wavelet analysis using continuous wavelet transformation and obtaining image of the wavelet coefficients in time-scale coordinate system for further processing (Howard et al. 2002). The following formula defines the coefficient calculation using time integral over the input signal using the wavelet:

$$C_{a,b} = \int_R s(t)\frac{1}{\sqrt{a}}\overline{\psi\left(\frac{t-b}{a}\right)}dt$$

where a – scale (frequency), b – time shift, C(a, b) wavelet coefficients.

Consider the following example earthquake (Figure 1): 2009-02-21, 01:29:01.1 GMT, 41.33 N 20.26E, Ml = 3.7, Albania. Recording stations: Tirana (Albania), Skopje, Bitola, Valandovo (R. Macedonia)

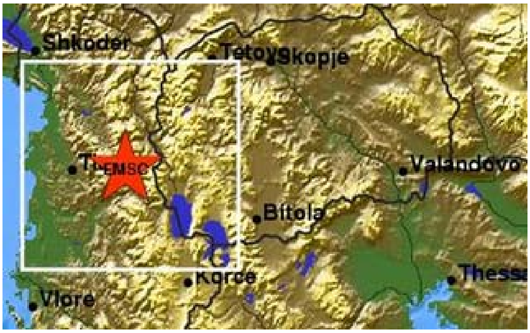


Fig. 1. Earthquake of 2009-02-21, 01:29:01.1 GMT, Albania

In our initial analysis, we have used the Mexican-hat wavelet. We plan on using different wavelets that have some physical meaning like Morlet wavelet. However it will be a subject of analysis and future research.

In the Figures 2-5 we represent the recording of the same earthquake as recorded from four different recording stations, Tirana – Albania; Valandovo, Bitola, Skopje – Macedonia respectively. In each figure the top graph represents time history, i.e. signal received from the station. Second (middle) graph is squared time history i.e. energy of the signal, while third (bottom) figure is image of continuous wavelet coefficients of energy time history.

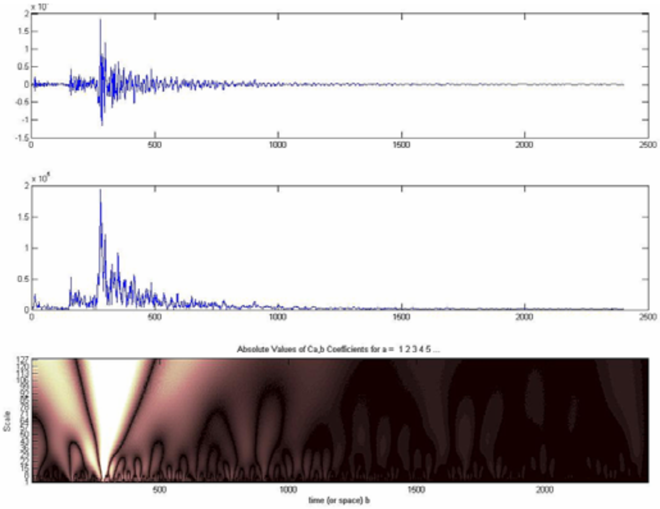


Fig. 2. Recording from station Tirana

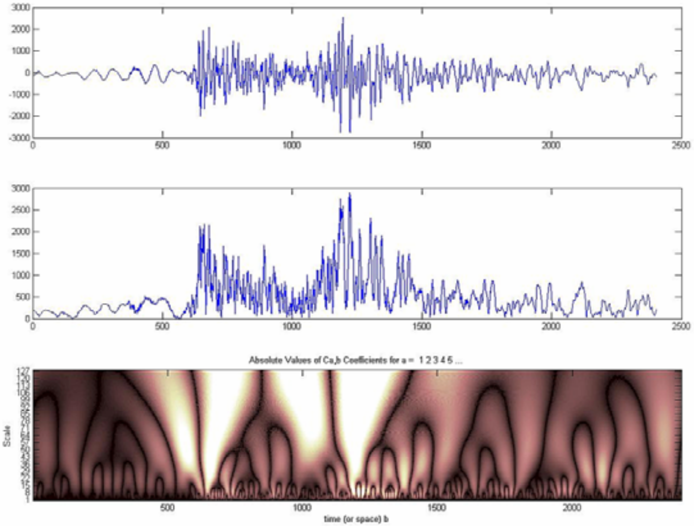


Fig. 3. Recording from station Valandovo

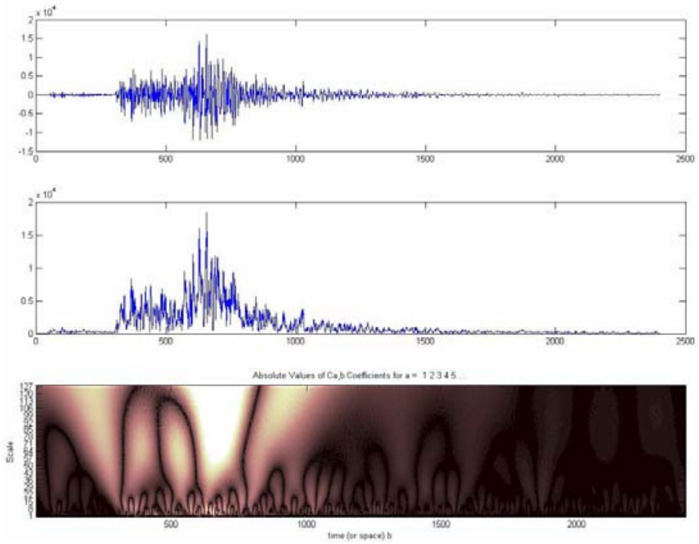


Fig. 4. Recording from station Bitola

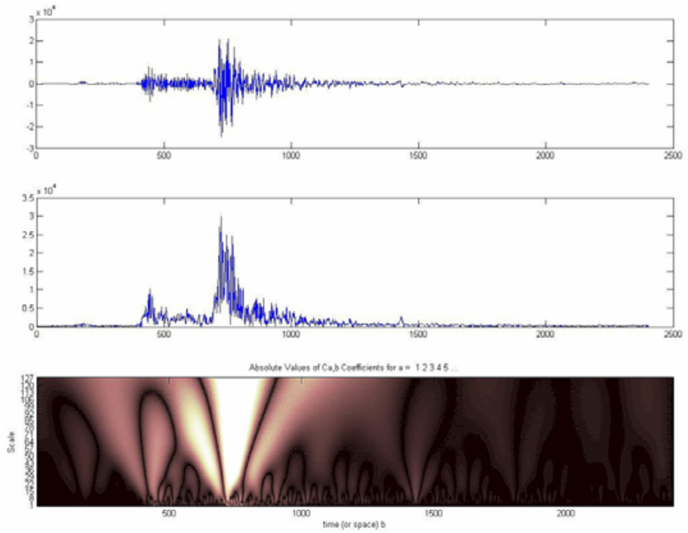


Fig. 5. Recording from station Skopje

In our example, the two stations Bitola (Figure 3) and Valandovo (Figure 4) have approximately same azimuth regarding epicenter. Bitola is on half distance between Valandovo station and epicenter. However due to same global geological profile between stations and hypocenter, images resemble each other. This is more pronounced on larger scale (longer period wavelet). Station Tirana (Figure 2) is closest to the epicenter and similarity of the images disappeared especially at larger scale (high period wavelets). Similar conclusion could be derived for station Skopje (Figure 5).

As mentioned in the examples there are similarities between the wavelet spectrums. In our initial research we noticed visual similarities. The spectrums were clearly influenced by many factors, which were captured by Continuous Wavelet Transformation. Since the comparison of images for pattern matching is very hard and inaccurate, we extracted features of the spectrum regarding the max lines occurrence, position and coefficient value. These extracted features are later processed by the pattern matching algorithm

Having analyzed a sufficient number of records of earthquakes for given station one can define similar patterns as a base for recognition of the epicentral zone and eventually to identify and describe geological path for given profile.

3 Application Model

The Grid application for processing the earthquake data was developed as a workflow application consisting of many separate grid jobs. The reason for this is that the processing of the seismic data can be logically separated into several phases, thus allowing for parallel and distributed processing of each of the phase. This kind of parallelization of the processing is best suited for the Grid environment, since it can concurrently use different Grid resources (i.e. different clusters).

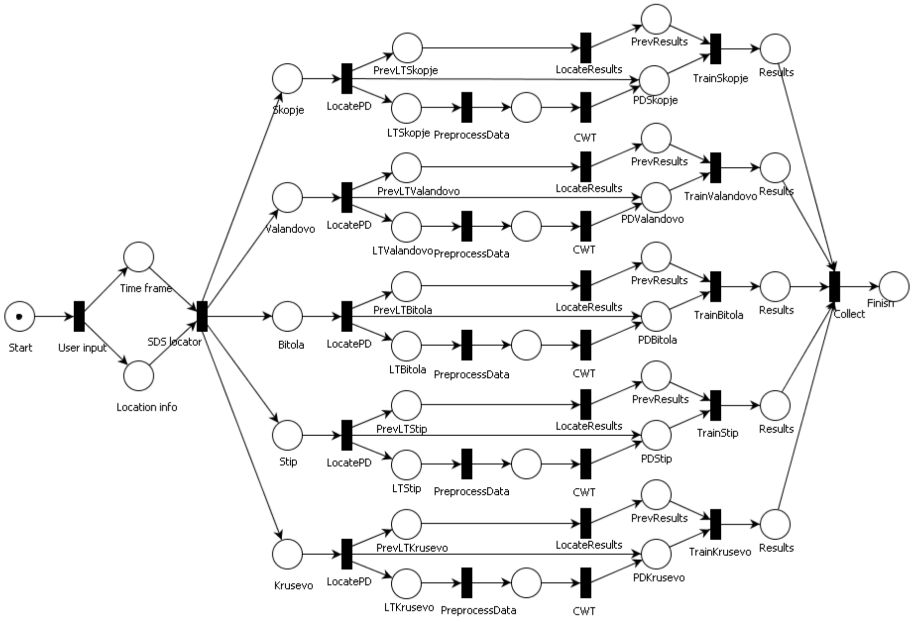


Fig. 6. Application Grid Workflow

In the course of the processing of the data, the Grid application uses several Grid services, available on the seismo.see-grid-sci.eu VO (deployed on the SEE-GRID infrastructure) in order to reach the intended purpose.

It first Grid job type (*SDS locator*) in the workflow is used for location of the data that will be processed. Using the already established Seismic Data Service (SDS) available on the seismo.see-grid-sci.eu VO one could iterate thru the available seismic data previously published on the Grid Storage Elements [10]. The SDS iterator allows for data search using the gLite AMGA service. When searching for data, one can search for seismic events that are available classified by regions. Second iterator allows searching of earthquake waveform data for a given time interval and station. These two iterators enable pinpointing the time intervals of earthquakes in a specified region, with specified magnitude limit and later finding Grid Logical Filename (LFN) of earthquake waveform data for specific stations.

Obtained time interval and LFN for each station are input data to the second Grid job type (*PreprocessData*). This information is used to contact the Logical File Catalog (LFC) and Storage Elements (SE) and retrieve the files containing the data for processing. Since the files can be in different formats like MiniSEED [8] and SAC [9], the next step is the conversion of the data for the specified time period from the specified station in the format appropriate for wavelet processing.

The third Grid job type (*CWT*) receives the preprocessed earthquake waveform data and performs the continuous wavelet analysis using the parameters specified in section 3. For the implementation of the CWT we used the open source Geophysical Wavelet Library [4] which we found very easy to work with and also very easily extensible with additional wavelet types.

The processing of the waveform produces earthquake wavelet spectrums that are stored on Grid Storage Elements and linked to LFC for later analysis. This is due the research needs for establishing good parameter extraction technique for better statistical classification. At the end of this Grid job the spectrum is analyzed and the characteristic parameters are extracted. The extracted parameters together with the information regarding the seismic region and receiving station are passed to the last Grid job type (*TrainStation*), classification training job.

The classification training Grid job (*TrainStation*) acts as a collector job of all processed data from one station. The extracted parameters, together with accompanying information regarding the earthquakes are fed as a training data to the pattern recognition algorithm in order to classify by epicentral regions the earthquakes that are received by that station. Since some of the data can already be available for that station, the classification training job can be preceded by a new job type – training data retriever that collects all available previously extracted parameters. The final results of the classification training are then stored for research purposes.

These four Grid job types are organized in several complex workflows depending on the need for the research. The most complete workflow contains all job types and is used for analysis of new earthquakes or new station/epicentral region pairs. The workflow is depicted in Figure 6 as a Petri-net containing all four job types. Parts of this workflow can be extracted as separate workflows that can be used for research issues while refining the parameters extraction and classification algorithm.

The workflow depicted in Figure 6 shows more than four mentioned Grid job types. The first transition (User input) contacts a web service to receive the new input data for the processing. This transition is a simple Web Service transition allowed in GWorkflowDL and is not a Grid job. After acquiring the user input, the second transition is the SDS Grid job specified earlier. After its completion SDS transition produces an output containing the LFN and time information of the seismic waveform. This information is given to the next transition – LocatePD that is also a Web Service transition. This web service contacts the web service of the application designed to keep all previously processed data. If the data was previously processed, the LocatePD transition will put a token in PD (Processed Data) and PrevLTSkopje places. If the data was not previously processed, the LocatePD transition will put a token in LT (LFC and Time) and PrevLTSkopje places. These transitions make possible for the workflow to be more efficient since it will not issue CWT processing of the data if it was previously processed. The token in LT place issues the next two consecutive Grid Jobs (*PreprocessData* and *CWT*) explained earlier and again at the end put a token in PD (Processed Data) place. The parallel transition LocateResults is again Web Service transition contacting the application web service for previously processed data for that station. Previously processed data supplied in the token located in the PrevResults place, together with the newly processed data located as a token in PD place are given to the last Grid job (*TrainStation*) for training the pattern matching algorithm. At the end the final transition, collect stores the final results on the Web Service of the application.

As it can be seen from the workflow only the first Grid job is stared only once, while the other Grid job types are started as a several concurrent jobs, one for each station analyzed. Current GWorkflowDL gLite subworkflow definition does not allow for asynchronous job execution, i.e. one transition will finish only when a job is successfully executed on the Grid. This limits the expressive power of the Petri-net workflow since it does not allow for arbitrary concurrent Grid job execution, i.e. the number of stations analyzed in parallel is static as opposed to a possible variable number depending on the available data. Further research in Grid Petri-net workflows needs to be done in order to overcome this limitation.

Execution of the defined GworkflowDL was using the Advanced Workflow Tool defined on the basis of the CPPWfMS [6] which is a Workflow Management System for the gLite middleware. The WfMS enables execution of complex workflows and is still under development. Having real problems while working on the developing WfMS shows its weaknesses and insures better final quality.

4 Conclusion and Future Work

In this paper we have presented a Grid framework for analysis and research of seismic waveform data by utilizing wavelets. The framework is based on several independent job types, each targeted to solve different parts of the problem. These jobs are later integrated into a complex workflow that allows automatic data acquisition using established SDS Grid service and its processing in order to produce the final results that are classified using pattern recognition training algorithms.

Next phase of the establishing friendly user environment for wavelet processing of seismic signal requires building large volume data bank of continuous wavelet spectrograms / extracted parameters to enable better understanding of the geological profile and its influence over wavelet spectrum.

The development of this application also opened many interesting research issues regarding the Grid Workflow execution. The gLite Grid sub-workflow model available in GWorkflowDL should be further investigated and opened to other kinds of execution in order to enable more flexible workflow design.

Acknowledgement

The work described in this paper is supported by the SEE-GRID-SCI FP7 project.

References

1. Aalst, W.: The Application of Petri Nets to Workflow Management. The Journal of Circuits, Systems and Computers 8(1), 21–66 (1998)
2. Aki, K., Richards, P.G.: Quantitative Seismology. University Science Books (2002)
3. Howard, L., Resnikoff, Raymond, O.W.: Wavelet Analysis: The Scalable Structure of Information. Springer, Heidelberg (2002)
4. Kulesh, M., Holschneider, M., Diallo, M.S.: Geophysical wavelet library: applications of the continuous wavelet transform to the polarization and dispersion analysis of signals. Computers and Geosciences 34(12), 1732–1752 (2008)
5. Mallat, S.: A Wavelet Tour of Signal Processing, 2nd edn. Academic Press, London (1999)
6. Pellegrini, S., Giacomini, F.: Design of a Petri Net-Based Workflow Engine. In: Proceedings of the 3rd International Conference on Grid and Pervasive Computing Workshops 2008, pp. 81–86 (2008)
7. Rao, R.M., Bopardikar, A.S.: Wavelet Transforms: Introduction to Theory & Applications. Prentice Hall PTR, Englewood Cliffs (1998)
8. IRIS Consortium, Standard for the Exchange of Earthquake Data (SEED) – Reference Manual. Verison 2.4, International Federation of Digital Seismograph Networks Incorporated Research Institutions for Seismology United States Geological Survey (2009)
9. Goldstein, P., Dodge, D., Firpo, M., Minner, L.: SAC2000: Signal processing and analysis tools for seismologists and engineers. In: Lee, W.H.K., Kanamori, H., Jennings, P.C., Kisslinger, C. (eds.) Invited contribution to The IASPEI International Handbook of Earthquake and Engineering Seismology. Academic Press, London (2003)
10. Balkır, A.S., Şenay, E., Unat, D., Özturan, C., Yılmazer, M.: Boğaziçi University: Kandilli Earthquake Seismic Data Server on Grid. Second Degree Workshop, Pula, Croatia (2007)
11. Hoheisel, A., Der, U.: An XML-based Framework for Loosely Coupled Applications on Grid Environments. In: Sloot, P.M.A., Abramson, D., Bogdanov, A.V., Gorbachev, Y.E., Dongarra, J., Zomaya, A.Y. (eds.) ICCS 2003. LNCS, vol. 2657, pp. 245–254. Springer, Heidelberg (2003)

Grids in the Near Future: A Technical and Social Review

Margita Kon-Popovska and Anastas Misev

University Sts Cyril and Methodius, Faculty of Natural Sciences & Mathematics
Institute of Informatics, Skopje, Macedonia
margita@ii.edu.mk, anastas@ii.edu.mk

Abstract. Crossing the institutional and national boundaries in the collaborative research has become reality. One of the most important tool enabling scientists to work on large projects, with massive parallelism and with huge datasets is of course the computing Grid. Although the technology has reached a level where big groups of scientists are using it on a daily basis, there are still many open issues and problems. From the technical aspects, there is still need for interoperability between many different middlewares, more intuitive and user friendly interfaces, tools to support richer workflows, Quality of Service tools and enablers, monitoring, etc. But there are also many other open issues, mainly related to the self sustainability of the infrastructure. Most of the current Grid infrastructure is supported by short term research and development projects. How will the infrastructure support itself is something to be seen in the near future. EU is strongly devoted to keep its leading position in the field of computing Grids, mainly focusing on integrating all the national Grid initiatives into a larger community, European Grid Initiative, EGI.

1 Introduction

The computing grids have become the foundation of the modern collaborative research, based on massive parallelism, huge datasets and expensive sensors. With the high speed networking below them, and the specific research projects and applications above, they are de facto the connecting layer that opens whole new research possibilities, unimaginable or unfeasible before. Recent development in the technology, especially in the grid middleware, raised the usability level of the computing grids to a reliable platform used by scientist in diverse areas on daily basis. Still, many issues still exits, issues that will make the grid infrastructure more appealing to the broader group of users. Inter grid cooperation, intuitive and user friendly interfaces, richer workflow support, QoS tools and enablers, richer monitoring tools etc. are only the few of the technical issues needed to be addressed in the near future. But the issues and uncertainties are not only in the technology itself. One general open question is the sustainability of the grid infrastructures. Currently, most of the operational infrastructures are built in the framework of various research projects with limited duration. How will this infrastructures support themselves after the project funding ends is to be seen in the near future. There is a strong devotion to support these infrastructures by the EU, thus maintaining its leading role worldwide in the area of computing grids. The most recent effort is the integration of all the national Grid initiatives into a larger community, European Grid Initiative - EGI.

In the first chapter we present some of the possible development directions for the grids in the near future. The second chapter contains information about the technical issues that will be in the focus of the future grid development. The next chapter discusses the social and political vision of the future grids. At the end, we conclude our review and present the used references.

2 Road from Here

There are several different views on what we expect from the grids. We will present some of them in this chapter.

One of visions for the grid is the community grids. There are several such initiatives, with the World Community Grid [1] being maybe the most widespread. It is an effort to create the world's largest public computing grid to tackle scientific research projects that benefit humanity. Launched in 2004, it is based on an IBM platform, with client applications for most popular operation systems. Once the client is installed on a users' computer that is internet connected, it will activate itself when the computer is idle. It will download a workunit from the central location, and when the work on that unit is completed, it will send back the results. Unlike other similar projects based on “cycle stealing”, focused on one problem only (Seti@home [2], Folding@home [3], ...), WGC supports wide variety of problems. Projects are approved by an advisory board, with members from many major research institutions and universities, as well as the U.S. federal government and World Health Organization.

Another quite different approach is presented in the work of Ganza et al [4]. At the foundation of their proposals is an agent based infrastructure. Agent, used with ontologies, infuses the grid with some form of intelligences M. Dominiak et al [5]. The resource brokering is based on sound business rules and Service Level Agreements - SLA. The authors consider their approach as returning to the basic idea of commercial exchange of resources (computing, storage), yellow pages approach for resource discovery, bargaining and agreement based on previous experience and current market offer, and strong mechanism to enforce compliance to the agreed SLAs.

The third approach is the cloud computing. It is a commercial, very restrained form of grid computing. The basic idea is to remove the computing out of the boundaries of the companies, into the area which is usually represented in the network diagrams as “the cloud”. In its current implementations (Amazon EC2 [22], Google App Engine[23], etc), it is a form of computing on demand. It is primarily based on the virtualization and provisioning of the computing and storage resources. Most of the implementation lack the basic elements of the grid computing, such as support for VO, workflows, and it basically offers “user-to-service” relation over public networks. The cloud computing is often referred to as Software as a Service, Utility computing, etc [21].

3 Technical Issues

According to a study called SUPER [6], to make the Grid infrastructure more useful for the scientists, three areas are identified: software, policy and support. In the software area, among other requirements, there is a need for “Better tools for understanding failures and increasing reliability”. Also, in the software area, the development of more intuitive user interfaces is strongly encouraged. Of equal importance is the policy area, where the recommendations are mainly focused on best practices, roles and relationships to support the virtual organizations. Finally, in the support area, more users and developer support is recommended, along with better training materials and demonstrations for self-learning.

One of the most important technical problems is the grid interoperability. At the moment, there are several popular grid middlewares (Globus [7], UNICORE[8] ….), that have very low or no interoperability at all. A way toward solving these issues might be the introduction of the Service Oriented Architectures – SOA, as envisioned in the Open Grid Service Architecture – OGSA [9]. When focusing on the interoperability, several communities are already working on joining the various grids into a single worldwide grid. There are some EU efforts [10], but one of the biggest initiatives is maybe the InterGrid [11], supported by Australian government and University of Melbourne. In their work, Marcos Dias de Assunção et al [12], identified many issues that have to be coped with to have lager interoperability. They identify the need for internetworking the current grid islands, establishing the analogy with the internet and the World Wide Web.

There are some limitations of the current middlewares that must be addressed and resolved. Important limitations are posed by the concept of virtual organizations. Currently, VOs are centrally managed and there is no possibility of dynamic VO establishment. Also, dynamics in the resource allocation per VO could contribute to better resource utilization.

Most important requirements for the grids of the tomorrow can be synthesized in the following major groups:

- Transparency and reliability, making the infrastructure available everywhere and anytime, so it can be used as a tool for daily operations.
- Openness toward wider groups of users, but also resource providers.
- Safety and security, based on mutual trust between the resource providers and consumers, regulated with standardizes security mechanism and policies.
- Ease of use and ease of programming.
- Interoperability between different resource allocation tools and policies, enabling intergrid collaboration at higher level
- Persistency, meaning that the users should rely on the platform to be useful not in the limited timeframe of the financing projects. Also, meaning that the infrastructure will evolve and accommodate emerging technologies.
- Scalability, enabling the infrastructure to grow and expand without any consequences to the currently established one.
- Standardizations of the protocols and software, with focus toward the OGSA.
- Ease of configuration and management. Self or auto configurability, similar to the next generation of networking protocols (IPv6 [13]).
- Clear and measurable QoS metrics, based on pricing of the resources and estimation of the requirements. This will enable sustainability, but also higher level of usage via more accurate scheduling and resource allocation.

Although some of the requirements mentioned are addressed by the current middlewares, most of them are still to be solved.

4 Social and Policy Issues

Several points from the social point of view need to be addressed to gain wider grid acceptance and usage. Among them, most important are:

- Users and institutions acceptance of the technology, especially the apparent loss of the "ownership" of the resources they contribute to the grid. This issue is usually countered with the fact that while they contribute their resources, they also gain access to much more resources.
- The acceptance of the idea of collaborative research as the basic research model when grids are used. Although for many of them, this will be a great possibility to work with larger teams, which was previously impossible, there are still scientists that believe that the grid will "steal" their ideas and jeopardize their individual results and involvement.

One of the biggest problems, already mentioned in the introduction, will be the sustainability of such infrastructures. Looking back, almost the entire existing grid installations are financed from various (mostly research) projects. As examples we could mention EGEE, SEE-GRID and many others, all within projects with limited duration. What will happen after the project funding stops is yet to be seen. There are lots of activities in this field, with the sole goal to help these infrastructures become self sustainable. The experience gained from the transitions of the research and education networks from the project funding toward self sustainability is of great help. In fact, the same model that was used for networking is being applied to the grids now. The European Grid Initiative – EGI [14] is the biggest pan-European activity to help this transition. As it was with the networks before, where each country established a NREN, that could later participate in joint projects and other forms of collaboration, the EGI suggests establishment of National Grid Initiatives. These NGIs will then join their effort in the EGI, as shown in Figure 1. The financing model for this national initiatives is based on several sources, including state investment into the NGI, joint European projects and finally through selling services to the industry. But such form of association poses whole new set of problems and challenges, including accounting, security, licensing, QoS, etc. D. Kranzlmüller [18].

The process of integrating the NGIs begun back in the 2005, when an advisory committee to the EU, called e-Infrastructure Reflection Group [15], identified that project funding of the grids is the major obstacle in their future. Following this recommendation, the idea of the EGI was constructed, presented in the Vision Paper. It was then circulated between the NGIs to obtain their support and to identify their representatives, forming the EGI Policy Board.

Members of EGI proposed an FP7 project called EGI Design Study (EGI_DS) with the EU's 7th Framework Programme, which is exploited to drive the developments of EGI by drafting corresponding documents and obtaining consensus within the NGIs. The EGI identifies some challenges that need to be addressed with corresponding actions, including:

- scalability, using federated and hierarchical approach, enabling wider collaboration;
- manageability on every level, starting from logical manageability of the VOs, toward mechanisms to enforce and implement SLAs;

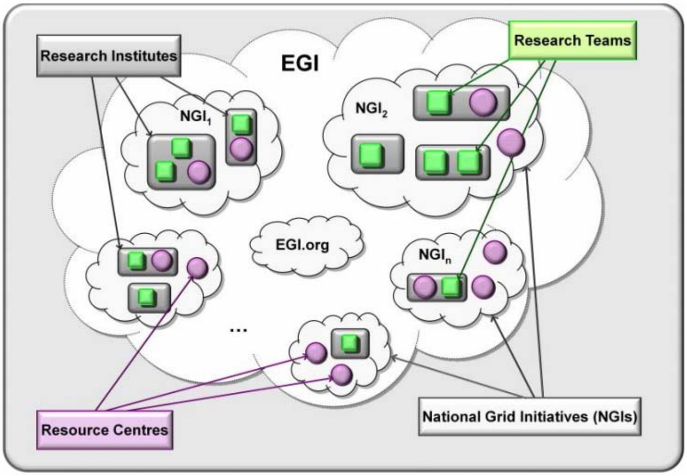


Fig. 1. EGI architecture proposed in EGI Blueprint [16]

- reliability, making the platform suitable for uninterrupted daily usage;
- interoperability and integration, enabling different middlewares, platforms and schedulers to seamlessly operate and exchange workload;
- higher level of services that will enable new ways of exploiting the infrastructure through their orchestration; and
- accounting and billing, necessary to provide business case to make the infrastructure sustainable.

The EGI management structure is shown in Figure 2.

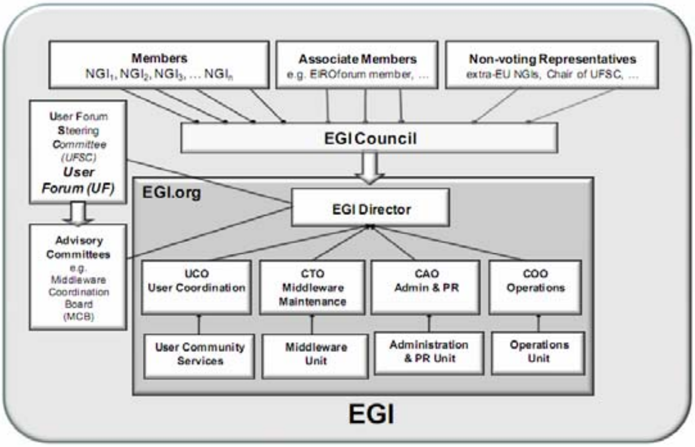

Members
NGI$_1$, NGI$_2$, NGI$_3$, ... NGI$_n$
Associate Members
e.g. EIROforum member, ...
Non-voting Representatives
extra-EU NGIs, Chair of UFSC, ...
User Forum Steering Committee (UFSC)
User Forum (UF)
EGI Council
EGI.org
EGI Director
Advisory Committees e.g. Middleware Coordination Board (MCB)
UCO User Coordination
CTO Middleware Maintenance
CAO Admin & PR
COO Operations
User Community Services
Middleware Unit
Administration & PR Unit
Operations Unit
EGI

Fig. 2. EGI management structure proposed in EGI Blueprint [16]

Europe is not the only one investing and basing its scientific potential on the grids. Open science grid [17] is the US equivalent of transcontinental grid infrastructure. It is developed to support collaboration in data intensive research by providing a computing facility and services that integrate distributed, reliable and shared resources to support computation at all scales, using facilities as shown in Figure 3. It supports multiple levels of activities toward the grater grid acceptance in US and wider. Most important activities are: the engagement of the potential new users, unfamiliar with the technology, assistance to merge the current university cluster infrastructures into the grid, education of both users and developers, interoperability toward other platforms and installations, even support for individual researchers and their applications.

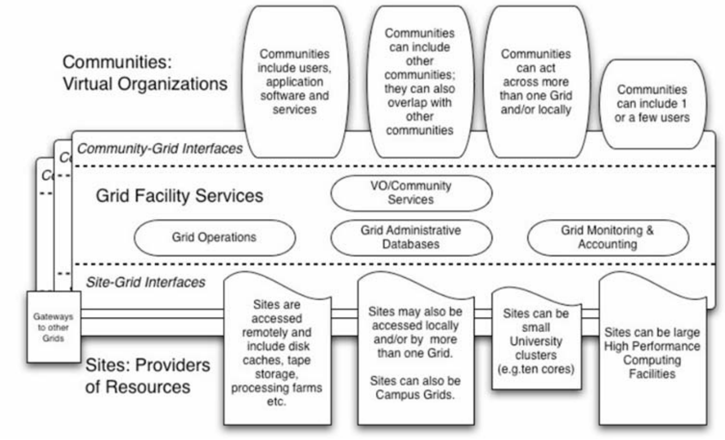

Communities: Virtual Organizations
Communities include users, application software and services
Communities can include other communities; they can also overlap with other communities
Communities can act across more than one Grid and/or locally
Communities can include 1 or a few users
Community-Grid Interfaces
Grid Facility Services
VO/Community Services
Grid Operations
Grid Administrative Databases
Grid Monitoring & Accounting
Site-Grid Interfaces
Gateways to other Grids
Sites: Providers of Resources
Sites are accessed remotely and include disk caches, tape storage, processing farms etc.
Sites may also be accessed locally and/or by more than one Grid.
Sites can also be Campus Grids.
Sites can be small University clusters (e.g.ten cores)
Sites can be large High Performance Computing Facilities

Fig. 3. Open Science Grid facility [17].

Perhaps the most important social issue is the users' perception of the grids. According to Liu[19], there are three aspects of the users' perspective: semantic, pragmatic and social. While the semantic view has been addressed well in previous work I. Foster at al [20], there is still lot to be done regarding the later two. The pragmatic aspect is about the understanding the context of the users and the support for their collaboration. It covers the modeling of the roles and access rights based on their responsibilities, duties and obligations. Such an approach helps defining the Virtual Organization – VOs and eases the collaboration through the grids.

One aspect that must not be neglected is the social aspect of the users' perception. It references the social, organizational and cultural rules for grid usage. Heterogeneity of the participating institutions, countries and people forming the VOs brings along heterogeneity in the working policies, rules and habits. Integration of such heterogeneity requires mutual understanding, tolerance and compromises.

5 Conclusion

Although the grid has become the computation platform for many scientists around the world, there are still many open issues that need to be addressed, so it will become more widely accepted and used. Issues involving QoS, resource pricing and utilization, security, dynamic resource allocation and discovery, fault tolerance, are only some of the aspects that need to be unraveled to have business involvement of the larger companies with the grid technology.

Many efforts are done not only from the technical, but also from the organizational and social point of view to support the involvement of the grids in new scientific areas, but also in the daily business. EU is investing most of its research resources toward the grid technology, trying to maintain its leading role in the usage of grids for science.

The ultimate goal of almost all worldwide grid research and development is the integration of all different grid middlewares and installations into a single, worldwide platform that will incorporate wide variety of processing, storage and sensory equipment, but also the know-how and user support, so that this platform can become the foundation for the future e-Science, much like the WWW was 15 years ago.

Acknowledgement

The work described in this paper is supported by the SEE-GRID-SCI FP7 project.

References

1. Word Community Grid (2009), http://www.eu-egee.org/
2. SETI@home (2009), http://setiathome.ssl.berkeley.edu/
3. Folding@home (2009), http://folding.stanford.edu/
4. Ganzha, M., Paprzycki, M., Drozdowicz, M., Senobari, M., Lirkov, I., Ivanovska, S., Olejnik, R., Telegin, P.: Information Flow and Mirroring in an Agent-Based Grid Resource Brokering System. In: Large Scale Scientific Computing, Sozopol 2009. LNCS. Springer, Heidelberg (2009)
5. Dominiak, M., Ganzha, M., Gawinecki, M., Kuranowski, W., Paprzycki, M., Margenov, S., Lirkov, I.: Utilizing Agent Teams in Grid Resource Brokering. International Transactions on Systems Science and Applications 3(4), 296–306 (2008)
6. Newhouse, S., Schopf, J.M., Richards, A., Atkinson, M.: Study of User Priorities for e-Infrastructure for e Research (SUPER). In: Proceedings of the UK All Hands Meeting (September 2007)
7. The Globus Aliance (2009), http://www.globus.org/
8. UNICORE (2009), http://www.unicore.eu/
9. OGSA-WG: Defining the Grid: A Roadmap for OGSA$^{\text{TM}}$ Standards (2005), http://www.ogf.org/documents/GFD.53.pdf
10. Grid interoperability project (2004), http://www.grid-interoperability.eu/
11. InterGrid: Internetworking Islands of Grids (2009), http://www.gridbus.org/intergrid/
12. de Assuncao, M.D., Buyya, R., Venugopal, S.: InterGrid: A Case for Internetworking Islands of Grids. Concurrency and Computation: Practice and Experience (CCPE) 20(8), 997–1024 (2008)
13. IPv6: The Next Generation Internet (2009), http://www.ipv6.org/
14. European Grid Initiative – EGI (2009), http://www.eu-egi.eu/
15. e-Infrastructure Reflection Group (2009), http://www.e-irg.eu/

16. EGI Blueprint (2009), http://web.eu-egi.eu/documents/other/egi-blueprint/
17. Open science grid (2009), http://www.opensciencegrid.org/
18. Kranzlmüller, D.: The Future European Grid Infrastructure – Roadmap and Challenges. In: ITI 2009 conference, keynote speech, Cavtat, Croatia (2009)
19. Liu, K.: Incorporating Human Aspects into Grid Computing for Collaborative Work. Keynote at the ACM International Workshop on Grid Computing and e-Science, San Francisco, June 21 (2003)
20. Foster, I., Kesselman, C., Nick, J.M., Tuecke, S.: Grid Services for Distributed System Integration. CACM, 37–46 (June 2002)
21. Armbrust, M., Fox, A., Griffith, R., Joseph, A.D., Katz, R., Konwinski, A., Lee, G., Patterson, D., Rabkin, A., Stoica, I., Zaharia, M.: Above the Clouds: A Berkeley View of Cloud Computing. Technical report, UC Berkeley Reliable Adaptive Distributed Systems Laboratory (2009), http://www.eecs.berkeley.edu/Pubs/TechRpts/2009/EECS-2009-28.pdf
22. Amazon Elastic Compute Cloud EC2 (2009), http://aws.amazon.com/ec2/
23. Google App Engine (2009), http://code.google.com/appengine/

Object Oriented Approach in Computer Aided Process Planning

Valentina Gecevska[1], Franc Cus[2], and Uros Zuperl[2]

[1] Faculty of Mechanical Engineering, University Ss. Cyril and Methodius,
Karpos II bb, P.O. Box 464, 1000 Skopje, Macedonia
valeg@mf.edu.mk
[2] Faculty of Mechanical Engineering, University of Maribor,
Smetanova 17, 1100 Maribor, Slovenia

Abstract. Process planning is one of the key activities for product design and manufacturing. Impact of process plans on all phases of product design and manufacture requires high level of interaction of different activities and tight integration of them into coherent system. In this paper, an object-oriented knowledge representation approach is presented with module for parts modeling and module for generation of process plan. Description of machining process entities and their relationships with features, machines and tools are provided. The benefits of the proposed representation, which include connection with geometric model, reduced search space and alternative plan generation, are discussed. These new contributions provide for a new generation of computer aided process planning (CAPP) systems that can be adapted for various manufacturing systems and can be integrated with other computer integrated manufacturing (CIM) modules.

Keywords: Process Planning, Knowledge, Object-Oriented Programming.

1 Introduction

Computer Integrated Manufacturing development has focused for a long period of time in linking various automated activities within the enterprise. However, the complexity of manufacturing process itself and extended application of computer supported equipment has led toward identifying three mail phases in manufacturing integration [2,5]: (1) hardware and software integration, (2) application integration and (3) process and people integration. After several years in focusing on CAD/CAM integration, the research has moved toward the third phase, process integration. One of most important links for implementation of integrated manufacturing is process planning, the link between product design (CAD) and production planning and execution (CAM, MES). Process planning as one of the key activities for product design and manufacturing is developed in many research. There are numerous papers devoted to various process planning systems which achieve certain level of manufacturing planning integration. Early major CAD/CAPP integration works are

[2], [3] that provide the integration between CAD and CAPP systems and provide the actual machining on NC machine connected to the system. Recent research efforts are devoted to generation and evaluation of alternative process plans and to enlargement of manufacturing knowledge base [1,2,4]. Integration with other manufacturing planning functions and the issues of data and knowledge representation and integration framework has also received significant interest [5].

This paper addresses an issue of generating process information within the integration framework in order to modeling manufacturing systems. The paper is organized in two sections. The first section describes interactions between manufacturing planning functions and identifies the need for integration. The second section explains manufacturing process' model and its modeling by object-oriented programming approach, developed in our research.

2 Manufacturing System Interaction

The product development and manufacture involves several production management activities with a series of individual tasks that are to be completed in order to design and manufacture a product of a required quality. These tasks are usually carried out in a linear sequence, but very often the feedback is necessary from the subsequent task to the previous one. Many of these feedback loops are requests to modify the previous task's solution in order to generate a better solution in the subsequent one. This interlinking is what has become known as concurrent or simultaneous engineering.

2.1 Manufacturing Activities Model

Starting from analyzing set of tasks of process planning and other activities, it is possible to develop the model that shows interactions between process planning and them. The model of these interactions is shown in Figure 1, where each activity represents with circle, consists a set of tasks that are to be done in the product development. All of these activities are identified in manufacturing planning literature as activities required during the product development and manufacture. The classification shown in the Figure 1 represents a starting point for the use of this method in each individual factory. There are numerous tasks that require interactions between two or more activities. They are shown within overlapping circles of activities and represent integration links.

It is important to understand explained interactions shown on the Figure 1 in order to completely utilize engineering knowledge and expertise. Each of these activities needs specialists in the domain, while intersections need group work and they are suitable for applying concurrent engineering principles. The most important intersections from process planning perspective are: between design and process planning related to part family formation, between process planning and resource management related to manufacturing cell design and between process planning and scheduling related to production control of cells.

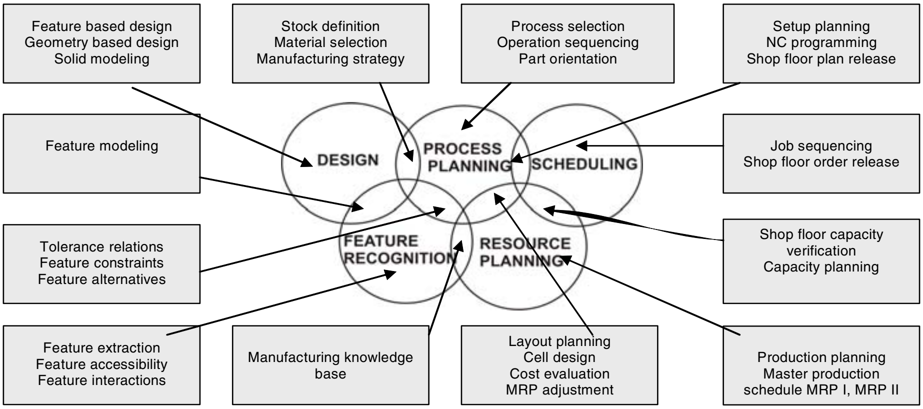


Fig. 1. Product development tasks

2.2 Process Planning Network

The basis of the above-described modeling of manufacturing process is the process planning network. This network is result of process planning that enable manufacturing integration. The process planning network consists of four interconnected layers: feature layer, process layer, tool orientation layer and machine layer. The first layer, feature layer, represents a network of machining features. The next layer is the process layer, which contains process candidate instances for individual features. In this layer it is necessary to show alternative machining processes for the same features in order to allow for later selection of the most suitable processes for given conditions. The next layer is tool condition layer that nodes represent sets of cutting process instances performed using tool orientation and machine type. The final layer of process planning network is machine layer. Its nodes set of cutting processes but with all constraints of machining conditions in order to determine - cutting data.

3 Object Oriented System for Process Planning

The process planning with several incorporated procedures described in this paper is part of a largest CAPP decision support system called OPTICAPP (Optimized Computer Aided Process Planning) [5], the general layout of which is shown in Fig.2.

The OPTICAPP system is implemented under the Microsoft Windows environment using the object-oriented programming with C++ program language. This system is structured with two basic program's module. Using a theory of graphs, mathematical logic and semantics, production rules and procedures are defined. Their structure is based on the interaction between technological orders of the machining (construction of workpiece, ordering of machining operations - mulltipass machining operations where one pass presented the sub-operation (rough or finish pass, or one cutting tool pass), priorities of the cutting operations, selection of tools, generation of optimal cutting data etc.). The optimization is made by a set of technological constraints, which include tool life, surface finish, machine power and available spindle speeds and feeds.

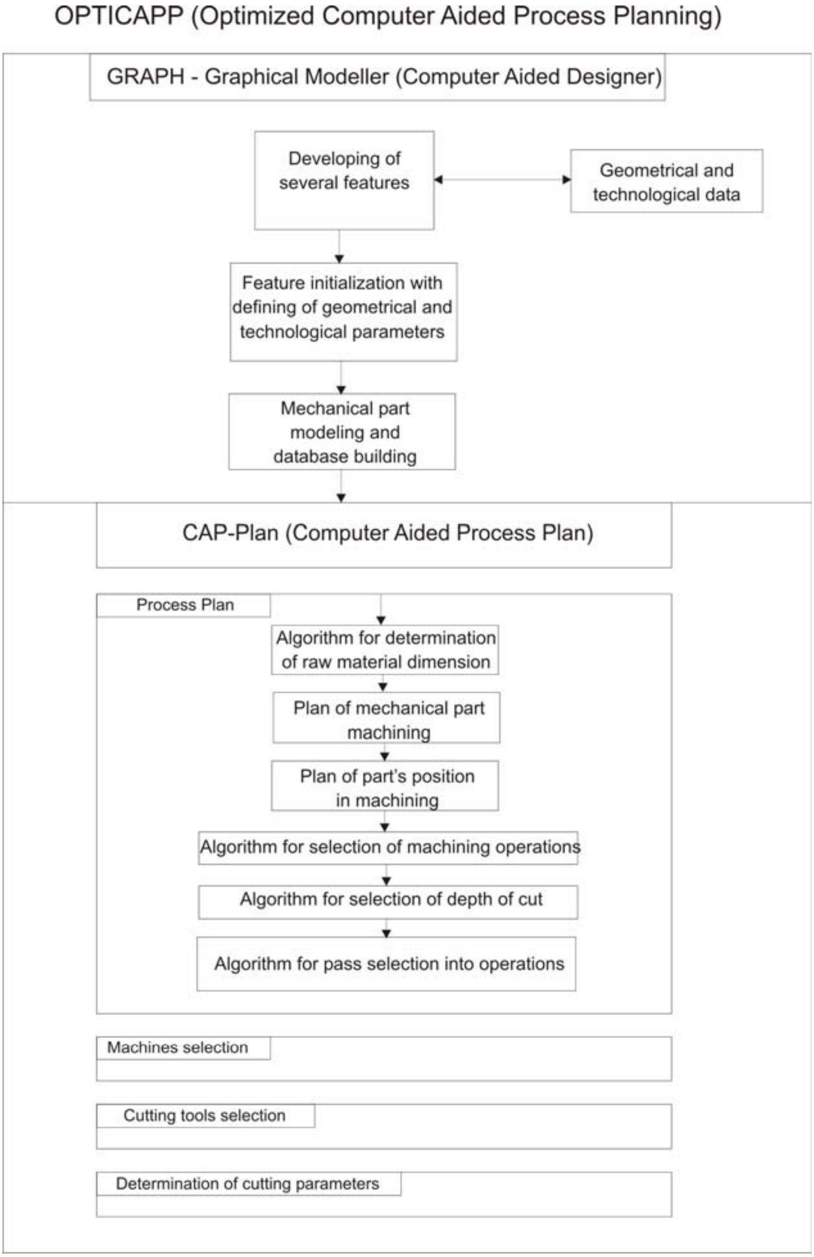

OPTICAPP (Optimized Computer Aided Process Planning)
GRAPH - Graphical Modeller (Computer Aided Designer)
Developing of several features
Geometrical and technological data
Feature initialization with defining of geometrical and technological parameters
Mechanical part modeling and database building
CAP-Plan (Computer Aided Process Plan)
Process Plan
Algorithm for determination of raw material dimension
Plan of mechanical part machining
Plan of part's position in machining
Algorithm for selection of machining operations
Algorithm for selection of depth of cut
Algorithm for pass selection into operations
Machines selection
Cutting tools selection
Determination of cutting parameters

Fig. 2. Overall layout of OPTICAPP

3.1 Module of Parametric Modeling Designer

The first module is computer aided designer, named GRAPH (Graphic Modeler). This module is used for work-parts modeling, based of features geometry recognition generated interactively by the operator, using a simple icon-based interface. The GRAPH modeler is based of a parametric method of modeling using developed icon based interface, showed on the Figure 3. Each icon represents one feature as an elementary work-part geometrical form. It is geometrical defined more than thirty different features, described with specific geometrical parameters (attributes). During modeling of one work-part it is necessary to selecting and definition of several features in according to the construction model of the work-part. During definition of geometrical parameters for each selected feature, this program module needs from user to define the technological parameters, as a work-part material, surface roughness and tolerances.

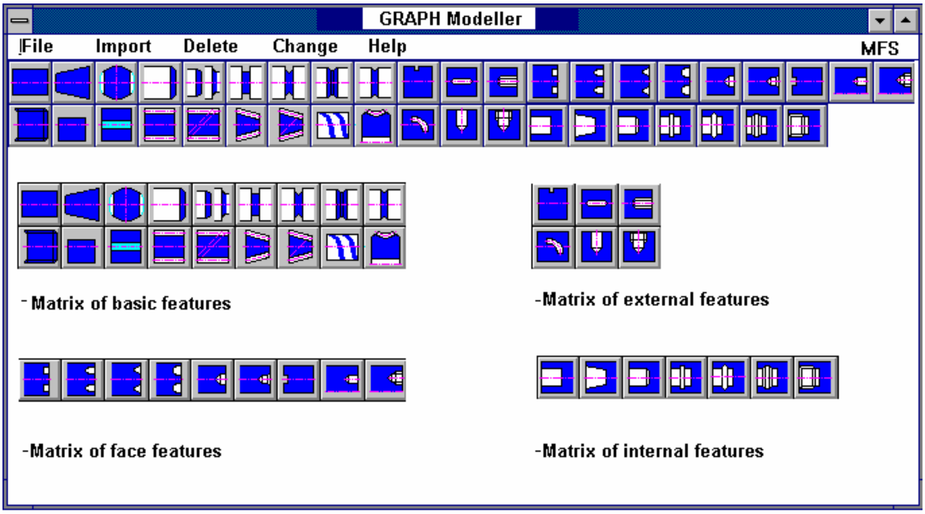

GRAPH Modeller
File
Import
Delete
Change
Help
MFS
- Matrix of basic features
-Matrix of external features
-Matrix of face features
-Matrix of internal features

Fig. 3. Developed icon based interface in GRAPH modeler

This module is based of the original developed mathematical methodology for modeling of mechanical part as a complex structure of features based of theory of graph, mathematical logic and semantics. The main characteristic of this module is the fact that, as a output, a dynamic model of the mechanical part is generated. The GRAPH modeler, for each modeled part, develops original dynamic database as a Double Linked List (DLL), core of the algorithm, holder of all geometrical and technological data for part. Dynamic database is based on the oriented multy graphs and data linking with pointers in DLL (Figure 4). Dynamic database for each modeled mechanical part is declared with follow program structure (struct rot type):

```
struct  rot type {
                    int  Prim;
                    char  *data;
                    float Data [MAXNUMPAR+1];
                    struct rot_type *prior;
                    struct rot_type *next;
                    }    * first, *last, *cur_rec;
```

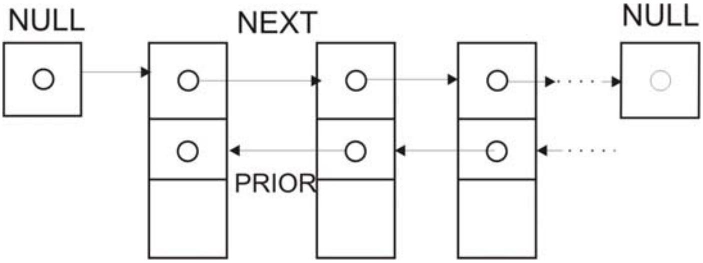


Fig. 4. Double Linked List (DLL) with graphical presentation of pointers (next / prior)

It is database for sorting and storing data for modeled mechanical part. Sorting of data in dynamic database is with network of knots and pointers (Figure 5), holder of data for all features in part's model.

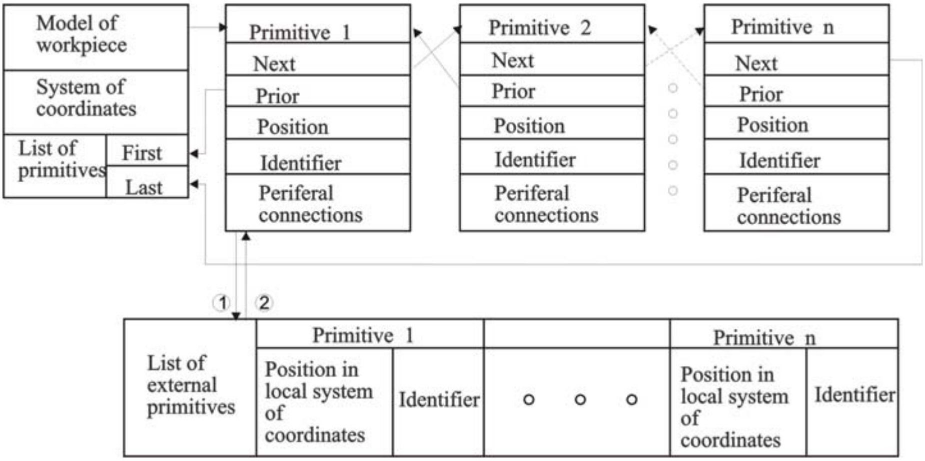


Fig. 5. Dynamic database with network of knots and pointers

The main advantage of this parametric modeling based design module is possibility for dynamical change of the created work-part model within the selected features. Software inter-communication is based of the complex part model (DLL with annexed oriented vectors) and many algorithms and procedures, as: algorithm for search and algorithm for subordination, here done with original code:

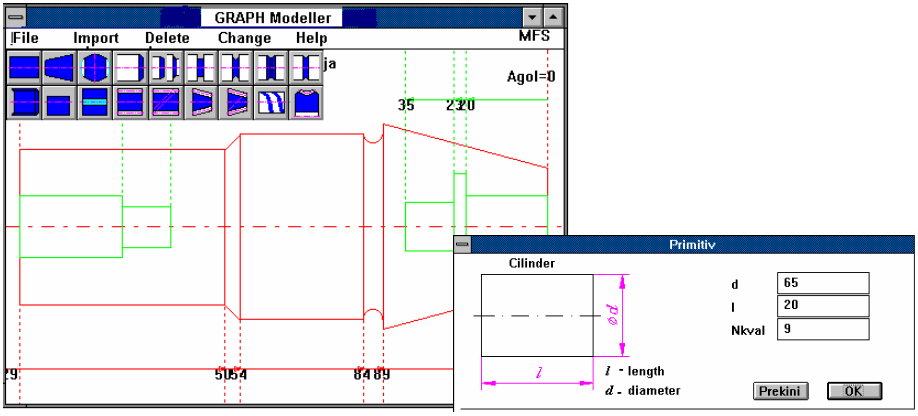
Algorithm for search

```
      info = from;
      if (way)     {
         while (info)  {
             num = (info->Prim);
         if     (     (key==(info->Data[0]))     &&
 character (num)=what )
                 return  info;
              info = info->next;   }
      else   {
          while (info)  {
             num = (info->Prim);
        if     (     (key==(info->Data[0]))     &&
 caracter(num)=what )
                 return  info;
              info = info->prior;   }
```

Algorithm for subordination

```
void left_face I ( void )
{
   struct rot_type *info, *temp;
   i=1;
   tem = info = first;
   Xorg=(MaxX+1)/40 ;   Yorg=MaxY/2;
   while ( info )   {
        info = find (temp, 1, CELEN);
              section I ( info, i, -1);
                                   from left to right
        temp = info->next;
        if (!info)  break;
          i++;          }
   return;  }
```

On the Figure 6, it is presented step by step (on going) modeling of mechanical part using features in GRAPH modeler.



d 65
l 20
Nkval 9

Fig. 6. Modeling of mechanical part using features in GRAPH modeler

3.2 Module for Generating Process Plan Network

The second module, it is generated process plan network for machining of modelled work-part by using the GRAPH modeller, module of OPTICAPP system. This module is named CAP-Plan (Computer Aided Process Plan). In this research, it is programming object-oriented algorithm using C++, based on a generative method with intelligent approach, as an expert system shell. This algorithm constitutes mathematical methodology with many developed rules and procedures for modelling of process planning. This algorithm can perform the following functions:

- Creating technological knowledge database for modelled mechanical part using the data from designer module,
- Presenting the technological knowledge with created rules for decision making,

- Algorithm for modelling process plan network for production with selection and succession on the operations and sub-operations,
- Modelling of each operation with order of sub-operation and optimized cutting parameters for each sub-operation,
- Select of cutting tools and machines for complete manufacturing process,
- Output presentation of generated process plan network.

3.2.1 Characteristics of the Object-Oriented Modelling for Process Planning

Decision for created a module for object-oriented modelling of process planning approach for automatic ordering of operations and sub-operation in machining process, it is based of the following reasons: generative approach of modelling as a method for individual design and modelling, recognition and process planning for each created mechanical part, open possibility for connection and integration of developed system in complex integrated CAD/CAPP/CAM system.

The module for process planning CAP-Plan, as inputs, uses:

- The dynamical database oriented model of the mechanical part generated in graphic modeller-GRAPH.
- The technological database of the model, carrier of a technological knowledge for modelled part necessary for process planning / generated in this program's module in the research/.

3.2.2 Technological Knowledge Database

Technological knowledge, as a basic for process planning, is composed of data and information, which is spectrum of know-how (Figure 7). Generally, the technological knowledge can be to divide on:

- Basic technological knowledge
- Experienced technological knowledge

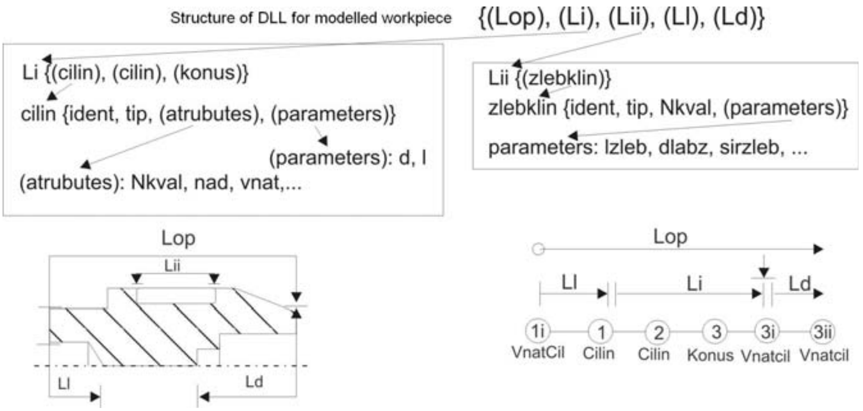

Structure of DLL for modelled workpiece
{(Lop), (Li), (Lii), (Ll), (Ld)}
Li {(cilin), (cilin), (konus)}
cilin {ident, tip, (atrubutes), (parameters)}
(parameters): d, l
(atrubutes): Nkval, nad, vnat,...
Lii {(zlebklin)}
zlebklin {ident, tip, Nkval, (parameters)}
parameters: lzleb, dlabz, sirzleb, ...
Lop
Lii
Ll
Ld
Lop
Ll
Li
Ld
1i
VnatCil
1
Cilin
2
Cilin
3
Konus
3i
Vnatcil
3ii
Vnatcil

Fig. 7. Logical structure of technological knowledge database

3.2.3 Mathematical Methodology for Technological Plan Design

Mathematical methodology for process planning and developing of technological plan is made with many algorithms and production's rules. The most suitable form for presentation of the technological knowledge is the form of modular production's rules, where each rule determines environs technological statement and it is free from other rule. Technological knowledge has been structured with the help of four types of the production's rules.

With object-oriented programming in the module for process planning, there are created following algorithms, which comprise logic for decision for each component of the process planning:

- The algorithm for define dimension on starting work part,
- The algorithm for design the form and the succession of the technological operation,
- The algorithm for modelling the actions into each operation (multi-pass operation),
- The algorithm for modelling the optimizing trajectory of cutting tool movement between the position points of machining.
- The algorithm for evaluation of justification for machining with one tool in all positions where it is predicted, or changing tool in certain position.

In the research, proposed OPTICAPP system, near two main modules for modeling and process planning, consists sub-module for optimization of cutting parameters for each machining operation and action, designed in process planning. It is developed as a multi-objective programming mathematical model where the optimal solution is obtained by using a deterministic method and a genetic algorithm, described in [8]. The cutting data optimization criterion is selected from minimum machining time or cost, maximum production rate.

3.2.4 Modelling the Action as Part of Technological Operation

The sub-operation, called action, is generated a surface on the mechanical part. For each surfaces a mechanical part, it is necessary projecting an action in the chosen operations. The requests that determine choice of the types of actions are the form, the dimension, the surface roughness, and the tolerances of the dimension and the position (Fig.8).

The actions are determined from the generated form. Respecting this aspect, the actions are analysed: (1) the actions are generated the form white repainting of the tool over the processing surface /ex. drilling/ and (2) the actions are generated the form white kinematics movement of the tools /milling, turning/.

Into processing the mechanical part, each action is separated the certain volume from materials of working part. New surfaces are generated in this action. In the research, for the process planning, it is applied a Backward Planning Method (BPM) (Figure 9), for described processing operations and processing actions, with added the layer of materials. For this method, starting position is finishing modelled mechanical part. With successive filling of layer materials, it is made the form of raw mechanical part. Each filling of the layer materials is presented the action or the operation on the processing.

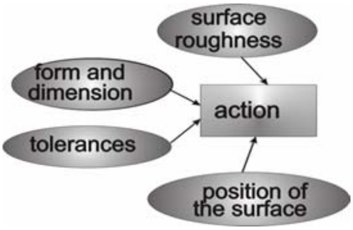


Fig. 8. The requests that determine choice of the type of action

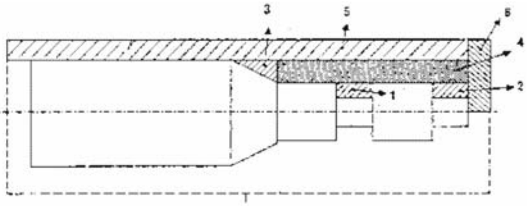


Fig. 9. Backward Planning Method

4 Conclusions

Manufacturing process model, as a basis for intelligent information integration, is based on analysis of interactions between various planning functions. The very important role of process planning function, as the function which defines manufacturing processes has been emphasized. It is proposed a knowledge representation scheme that recognized geometric and feature-based representation of parts should be considered in the design of process planning expert systems. Proposed programme system with two major modules done part representation for process planning that includes the general part description, the set of features with geometric and technological information, and the object-oriented programming with double linked list which describes the interactions among features.

The major contribution of this paper is its comprehensive knowledge representation system that includes the following approaches: (1) connection of feature and process knowledge with the part geometric model; (2) object-oriented technological database for presentation of technological knowledge and (3) generation of proposed process plans. These new contributions provide for a new generation of CAPP systems that can be adapted for various manufacturing systems and can be integrated with other CIM modules.

References

1. Carpenter, I., Maropoulos, P.: Automatic tool selection for milling operations Part 1: cutting data generation. Journal of Engineering Manufacture 214, Part B, 271–282 (2001)
2. Chang, T.: Expert Process Planning for Manufacturing. Addison-Wesley, Reading (1990)
3. Balic, J., Pahole, I.: Optimisation of intelligent FMS using the data flow matrix method. Journal of Materials Processing Technology 133(1/2), 13–20 (2003)
4. Sormaz, D.: Intelligent Manufacturing Based on Generation of Alternative Process Plans. In: Proceedings of 9th Int. Conference on Flexible Automation and Intelligent Manufacturing, Tilburg, pp. 35–49 (1999)
5. Gecevska, V., Cus, F., Kuzinovski, M., Zuperl, U.: Evolutionary Computing with Genetic Algorithm in Manufacturing Systems. Journal of Machine Engineering 5(3/4), 188–198 (2005)
6. Jain, A., Jain, P., Singh, P.: Deadlock Analysis in FMS in the Presence of Flexible Process Plans. International Journal of Simulation Modelling 4(2), 53–66 (2005)

Videoconferencing as Tool of Higher Education

Enrica Caporali[1], Vladimir Trajkovik[2], and Juna Valdiserri[1]

[1] School of Engineering, University of Florence, Via S. Marta, 3, 50139 Firenze, Italy
enrica.caporali@unifi.it, vices@unifi.it
[2] Faculty of Electrical Engineering and Information Technologies,
"Ss Cyril and Methodius" University, Skopje, R. Macedonia
trvlado@feit.ukim.edu.mk

Abstract. In the decade up to 2020 European Higher Education will have a vital contribution to realize a Europe of knowledge that has a relevant role, at national and international level, in the cultural and economical development of countries. The higher education will also face the major challenges and opportunities of globalization with accelerated technological developments with new providers, new learners and new types of learning. New educational requirements stimulated by the innovative telecommunication technologies, leads, almost as direct consequence, to the latest educational materials and methodologies, to videoconferencing and distance learning issues. In this framework, the three-year ViCES (Video Conferencing Educational Services) Project was launched and financed by the European Commission within the TEMPUS (Trans-European Mobility Scheme for University Studies) programme. The VICES project will provide an environment that increases student and academic mobility as well as infrastructure that will ease the process of harmonization of different curricula outcomes.

Keywords: higher education, TEMPUS European Programme, innovative educational methods, distance learning, curricula harmonization, national and international cooperation.

1 Introduction

Higher Education plays a very important role in the development of human beings and societies and enhances cultural and economical development as well as expertise for educational growth. It includes in fact teaching, research and social services activities of Universities and refers to education provided by Universities, colleges, institutes of technology and other collegiate level institutions awarding academic degrees or professional certifications [1].

Since 1950, article 2 of the first Protocol to the European Convention on human Rights [2] obliges all signatory parties to guarantee the right to education. World-wide, the United Nations' International Covenant on Economic, Social and Cultural Rights of 1966 [3] guarantees this right under its Article 13, which states that «higher education shall be made equally accessible to all, on the basis of capacity, by every appropriate means, and in particular by the progressive introduction of free education».

The European Higher Education Area is the objective of the Bologna Process [4][5], devoted, since 1999, to create more comparable, compatible and coherent education systems throughout Europe, based, among others, on the European Credit Transfer System – ECTS. The higher Education institutions, to achieve the objectives above and to encourage cooperation between countries, may take part in a wide range of programmes, such as LLP (Lifelong Learning Programme) [6], ERASMUS MUNDUS [7], TEMPUS (Trans-European Mobility Scheme for University Studies) [8].

The TEMPUS Programme is designed to support the "transition and modernization processes" in higher education through a range of interventions and to create an area of co-operation in countries surrounding the EU. Established in 1990 after the fall of the Berlin Wall, the scheme now covers 27 partner countries in the Western Balkans, Eastern Europe and Central Asia, North Africa and the Middle East [8]. In addition to "people to people" academic cooperation, Tempus aims at having an impact on higher education policies, and closely following national higher education priorities.

Higher education and learning are taking place across the whole life span in a wide range of environments and with different aims. New educational requirements and innovative education practices stimulated by the new information and telecommunication technologies enable all the actors involved in the educational process almost instantaneously to access the latest educational materials and methodologies. Students are usually familiar with the use of different technologies for their studies and research. This fact opens the possibility for creating an education environment, where high end internet based services are used to implement techniques which cannot be implemented in traditional classrooms. Technology itself is not inherently good or bad for educational process support. It is the way it is used that matters.

Video conferencing enabled learning is a new way of acquiring knowledge, which is highly adaptable to different kinds of student profiles, from people that do not have time to attend normal courses to a practical enhancement of ordinary courses with additional access to the knowledge. It facilitates and promotes the co-operation, at national and international level, generating new networks and more immediate communications processes of personal and professional contacts. Exchange of knowledge and consultation process among students and available expert authorities (professor/instructor), are very important aspects of learning, in addition to the static contents that are provided in books and different digital multimedia.

In this framework, the University of Florence and the Ss Cyril and Methodius University launched in 2008 a three-year TEMPUS JP Project called VICES (Videoconferencing Educational Services) financed by the European Commission in the frame of the TEMPUS IV for the period 2009-2012 [9].

The project, carried out by the University of Florence and the Ss Cyril and Methodius University in Skopje, together with all consortium members (three partner Universities of the European Union and different Universities in Albania (AL), Republic of Macedonia (MK) and Serbia (RS)), will introduce a new approach towards the treatment of Information Communication Technologies at University level. It is expected that VICES will provide an environment that supports and increases student and academic mobility as well as infrastructures that will ease the process of harmonization of different curricula among educational institutions.

2 Higher Education in Europe: From the Bologna Process to the Tempus Programme

One very important step to reach more comparable and more compatible educational systems in Europe can be initially identified in the Recognition of Qualification concerning Higher Education in the European Region which considers the great diversity of education systems in the European region and tries to identify a strategy for the recognition of studies, certificates and diplomas and degrees obtained in different countries of the European Area [11]. After that, the Bologna Process [4][5] (started in 1999 with the signature of the Bologna Declaration from the ministries of Education of different European countries) aims to create a European Higher Education Area (EHEA) based on international cooperation and academic exchange that is attractive to European students and staff as well as to students and staff from other parts of the world.

The EHEA aims to: facilitate mobility of students, graduates and higher education staff; prepare students for their future careers and for life as active citizens in democratic societies, and support their personal development; offer broad access to high-quality higher education, based on democratic principles and academic freedom. Some of the Bologna action lines are qualification frameworks (Three cycle System), joint degrees, mobility, recognition, quality assurance, social dimension, employability, lifelong learning.

The Bologna Process is taken forward through a work programme that receives orientations from ministerial conferences every two years (Praga 2001, Berlin 2003, Bergen 2005, London 2007, Leuven 2009) [4]. These conferences are prepared by a Bologna Follow-up Group, which in turn receives input from working groups and Bologna Seminars. The last ministerial conference in Leuven last April 2009 [4], stresses the achievement on the Bologna Process until 2009 and define the priorities in education on the decade up to 2020 in terms of Social dimension (equitable access and completion), lifelong learning, employability, student-centred learning and teaching mission of higher education, education, research and innovation, international openness, mobility, data collection, multidimensional transparency tools and funding. In the Leuven Communiqué the European Ministers responsible for Higher Education state that the Bologna Process is leading to greater compatibility and comparability of the systems of higher education and is making it easier for learners to be mobile and for institutions to attract students and scholars from other continents with a constant focus on quality.

Access to higher education is expected to be provided by fostering the potential of students from underrepresented groups. This involves improving of the learning environment, removing barriers to study, creating the appropriate economic conditions for students to be able to benefit from the study opportunities at all levels towards the achievement of equity in higher education. Lifelong learning is perceived as an integral part of the European education systems. Recommendations to assure the accessibility, quality of provision and transparency of information are included. Lifelong learning involves obtaining qualifications, extending knowledge and understanding, gaining new skills and competences or enriching personal growth.

Lifelong learning implies that qualifications may be obtained through flexible learning paths, including part-time studies, as well as work based routes.

Student-centred learning requires empowering individual learners, new approaches to teaching and learning, effective support and guidance structures and a curriculum focused more clearly on the learner in all three cycles. Curricular reforms are identified as an ongoing process leading to high quality, flexible and more individually tailored education paths.

It is expected that Higher education should be based at all levels on state of the art research and development thus fostering innovation and creativity in society. The potential of higher education programmes, including those based on applied science, to foster innovation is recognized. European higher education institutions are called to further internationalise their activities and to engage in global collaboration for sustainable development. The attractiveness and openness of European higher education is highlighted by joint European actions.

The mobility of students, early stage researchers and staff is considered as an added value in the quality of programmes and excellence in research and for the academic and cultural internationalization of European higher education. Mobility is important for personal development and employability; it fosters respect for diversity and a capacity to deal with other cultures. It encourages linguistic pluralism, underpinning the multilingual tradition of the European Higher Education Area and it increases cooperation and competition between higher education institutions.

The above mentioned priorities are taken into account in a number of initiatives of the European Commission and Member States in the field of education and training. The Council Conclusions on a strategic framework for European cooperation in education and training ("ET 2020"), adopted in May 2009, build on progress made under the previous work programme and set four strategic objectives: making lifelong learning and mobility a reality; improving the quality and efficiency of education and training; promoting equity, social cohesion and active citizenship; enhancing creativity and innovation, including entrepreneurship, at all levels of education and training. These activities also contribute to the Bologna intergovernmental process in the field of higher education.

One of the main European funded co-operation Programme in the field of Higher Education is the TEMPUS Programme [8] (Trans-European Mobility Scheme for University Studies) (Fig. 1). It supports the modernisation of higher education and creates an area of co-operation in countries surrounding the EU. Established in 1990 after the fall of the Berlin wall, the scheme now covers 27 countries in the western Balkans, Eastern Europe and Central Asia, North Africa and the Middle East (going beyond the «iron curtain»). It strengthens cooperation in higher education between the European Union and its partner countries and at the same time it enhances understanding between cultures, promoting the “people to people” approach.

The overall objective of Tempus is to contribute to the creation of an area of cooperation in the field of higher education between the European Union and Partner Countries in the countries neighbouring the EU. The specific objectives of Tempus are: to promote the reform and modernisation of higher education in the Partner Countries; to enhance the quality and relevance of higher education to the world of work and society in the Partner Countries; to increase the capacity of higher education institutions in the Partner Countries and the EU, in particular their capacity to co-operate internationally and to continually modernize, and to assist them in opening up to society at large, the world of work and the wider world in order: to overcome inter-country fragmentation in the area of higher education and inter-institutional fragmentation in countries themselves; to enhance inter-disciplinarity and trans-disciplinarity between university faculties; to enhance the employability of university graduates; to make the European Higher Education Area more visible and attractive in the world; to foster the reciprocal development of human resources; to enhance mutual understanding between peoples and cultures of the EU and the Partner Countries.

The TEMPUS Programme can finance two types of actions: Joint Projects, based on multilateral partnerships between higher education institutions in the EU and the partner countries. They can develop, modernise and disseminate new curricula, teaching methods or materials, boost a quality assurance culture, and modernise the management and governance of higher education institutions. The second type of action are Structural Measures which contribute to the development and reform of higher education institutions and systems in partner countries, enhance their quality and relevance, and increase their convergence with EU developments.

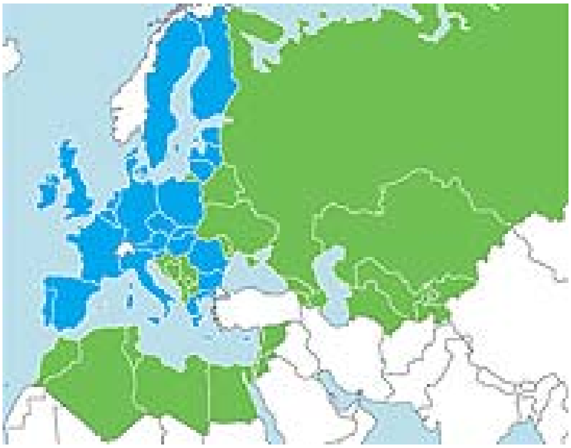


Fig. 1. The countries in blue are the European Union countries and the countries in green are the partner countries to which the TEMPUS Programme is addressed [8].

3 From the Video Conferencing Approach to the TEMPUS VICES Project

Comprehensive new approaches to valuable learning, which will allow citizens to move freely between learning settings, jobs and countries, making the most of their knowledge and competences, should be always considered as very important for every community [12] [13]. Videoconferencing services, used in combination with other educational services significantly ease this access by lowering the cost of original production of educational material and increasing the possibility to update educational materials more frequently.

Video conferencing involves a two-way video, audio and data communication between two or more parties over a remote connection [14]. Video conferencing is carried out over a variety of media, the most popular of which uses Internet Protocol (IP) technology. The cost of video conferencing over IP is getting so low that it has become the most popular means of video conferencing [15].

Streaming technology is considered to be a very important internet based network technology that enables the deployment of video conferencing services. Streaming technology covers one way transmission of audio, video and possibly other content to an end user. When speaking about videoconference, archiving and subsequent methods of retrieval of archive content must be specified. Real-time streaming of videoconferences is of great importance as a videoconference service. It should be noted that the quality of service of the real time video streaming is of great importance for the end user perception of the content. This importance is even increased in case of bidirectional interactive streaming such as video conferencing in education. The second technical issue important for video conferencing is the video format resolution [16]. Figure 2 compares the typical formats used in different video standard that can be streamed. It is obvious that more audience requires better video resolution. Unfortunately, this makes providing the needed quality of service more complex. That is the main reason why it is important to build large a scale video conferencing educational system over the manageable network infrastructure. A Typical example of such an infrastructure is the national academic internet network.

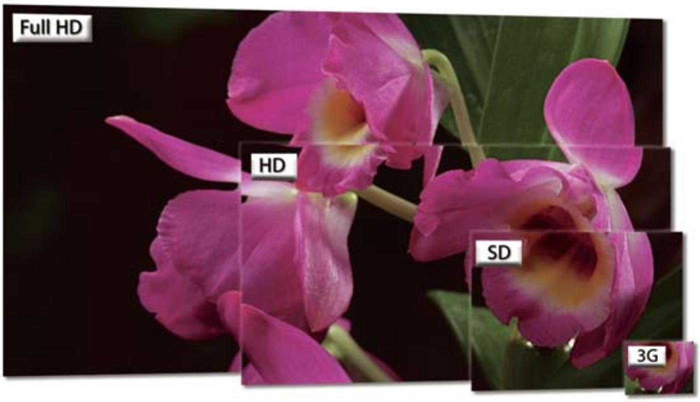


Fig. 2. Typical resolutions for standard video formats

The components of the learning environment that promote the usage of video conferencing services can be itemized as follow: educational methodology used in the learning process, mapping of video conferencing technology onto the educational methodology, and institutional factors influencing the educational process.

In order to make video conferencing to function effectively, the instruction and course content must be interactive, and the instructor must exhibit flexibility and creativity when teaching the class. At the same time, he/her has to able to manipulate multimodal content (video, audio, and data) that should be presented to the students. In addition to this, technical support for managing the video conferencing equipment is required. A typical scenario for the 30 students' classroom includes 2 video screens that are presented to the students at the same time – one for standard video conferencing and one for educational material. Figure 3 illustrates those video screams and related video sources. It is obvious that a certain technical knowledge has to be provided for such classroom.

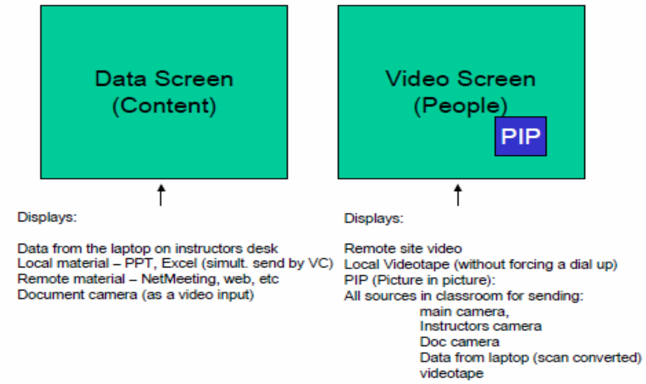

Data Screen (Content)
Video Screen (People)
PIP
Displays:
Data from the laptop on instructors desk
Local material – PPT, Excel (simult. send by VC)
Remote material – NetMeeting, web, etc
Document camera (as a video input)
Displays:
Remote site video
Local Videotape (without forcing a dial up)
PIP (Picture in picture):
All sources in classroom for sending:
main camera,
Instructors camera
Doc camera
Data from laptop (scan converted)
videotape

Fig. 3. Elements of Multimodal presentation in video conferencing classroom

Video conferencing enhanced distance learning increases educational opportunities offered by any institutions. It reduces the costs of teaching and learning, while allowing students to have more access to a variety of degree programmes. The management of video conferencing-based education is dependent on the geographical distance of the participants and the number of separate sites involved in the interaction.

Videoconferencing can help to make the different systems of higher education more compatible and comparable and to promote equal opportunities to quality education. Furthermore, it can guarantee accessibility, lifelong learning as an integral part of education systems by introducing flexible learning paths.

Furthermore, the use of innovation technologies in learning foster innovation and creativity in society, enabling students to be more employable and more advanced in knowledge, skills and competences. Videoconferencing can be considered as an important ICT tool to carry out the priorities of the Bologna Process in the next decade.

There are other factors which influence the successful implementation of educational processes. These factors relate to the institutional needs in higher education:

- The need for large scale collaboration in education technology development.
- The need to share resources, especially transferable courseware, on a national scale.
- The need for staff development.

The establishment of a video conferencing infrastructure and corresponding educational methodology (two outcomes of the TEMPUS VICES project) [9] will be the basis for further development of an efficient lifelong learning universities' educational system.

4 Expected Results and Future Perspectives of VICES

The Macedonian Universities have commitments to protect and preserve the cultural heritage of the region and its citizens. One of the most significant cultural heritages is the language. Thus, at the Universities in Macedonia, lectures are given in Macedonian, Albanian and English language depending on the needs of the students that attend different classes. In this context a multi-cultural and multi-lingual educational environment is created.

One of the main objectives of the TEMPUS VICES project (Videoconferencing Educational Services) [9] will be to enable the usage of sophisticated video conferencing and other distance learning environment services in combination with traditional face to face learning in order to establish the ratio of different learning methodologies most suitable for the students' needs taking into account cultural, technical and economical partner country background and needs [17].

The VICES project will provide one centred Video conference management system and seven video conference classrooms in R. Macedonia, as well as two video conferencing classrooms in Albania and Serbia. The general scheme of the VICES video conferencing infrastructure is given on Figure 4.

The Video Conference management centre will be facilitated by the Macedonian Academic and Research Network (MARNET), due to the already established management of the academic network infrastructure. This equipment will consist of three parts: management software, recording and streaming server and multipoint conference units. The management software will be able to utilize and optimize the network traffic generated by the video conferencing sessions. The recording and streaming server will provide recording capabilities for any video conferencing sessions, thus enabling their later streaming to any web enabled client [18]. It has to

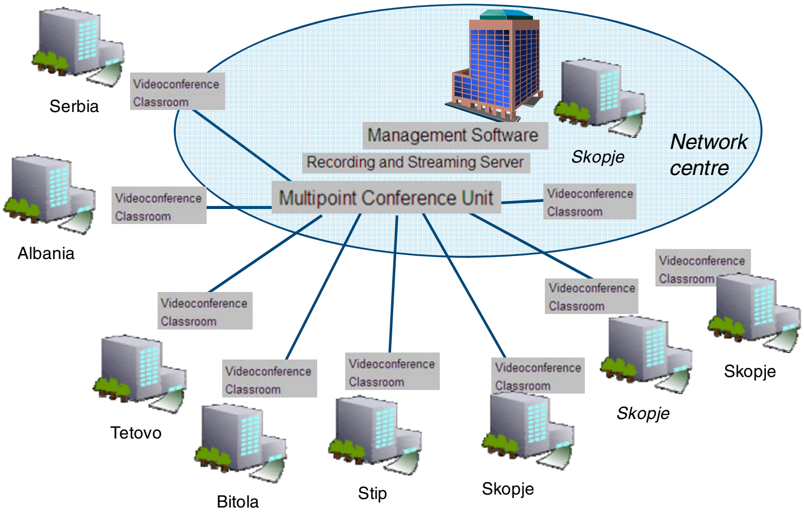
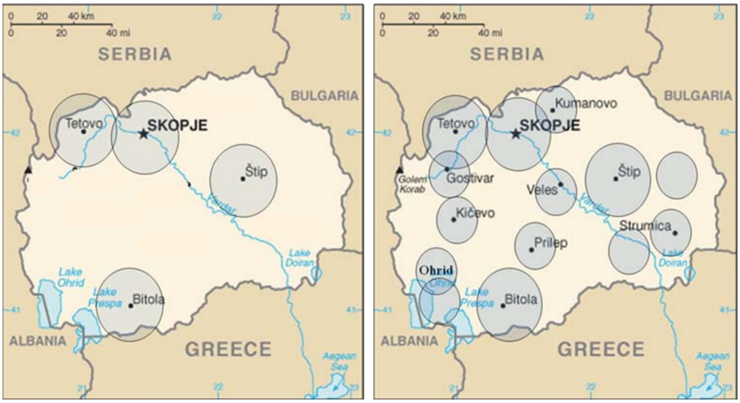

Serbia
Videoconference Classroom
Management Software
Recording and Streaming Server
Skopje
Network centre
Albania
Videoconference Classroom
Multipoint Conference Unit
Videoconference Classroom
Videoconference Classroom
Videoconference Classroom
Videoconference Classroom
Tetovo
Videoconference Classroom
Videoconference Classroom
Videoconference Classroom
Skopje
Skopje
Bitola
Stip
Skopje

Fig. 4. The VICES video conferencing infrastructure

be stated, that in this case, the students will not be able to interact with their instructors. The multipoint conference units should enable parallel and multicast session among different video conferencing classrooms. In this way, using the video portal provided by the VICES project, students from different Universities will be able to attend different lectures on the same or similar subjects. Students will be able to exchange their ideas and educational findings with wider student communities that share similar interest [19].

Figure 5a presents the initial geographical placement of the video conferencing classrooms within R. Macedonia covered by the VICES project, while Figure 5b presents the potential video conferencing classrooms. The potential classrooms locations are determined according the locations where the Universities that participate in the project have dispersive centers. As it can be seen from this figure, it covers significant population in R. Macedonia, providing equal access to all students to the higher education facilities.



(a)

(b)

Fig. 5. (a) Initial and (b) potential placement of video conferencing classrooms

The establishment of video conferencing infrastructure and corresponding educational methodology will be the basis for further development of an efficient lifelong learning universities' educational system.

The VICES Videoconferencing Educational Services will be evaluated by students at their last year of undergraduate studies in Information Technologies using standard evaluation techniques adopted to video conferencing systems [20][21]. The students will be asked whether the video conferencing is useful for their studies. The questionnaires given to the students for evaluation will include three types of questions regarding: student experience in using video conferencing technologies in education, multimodal accessibility of the educational content, and quality of service of video conferencing.

5 Conclusion

Higher Education will play a central role to realize the Europe of knowledge in the decade up to 2020. A great number of initiatives are actually carried out in this direction from the European Commission and member states.

One of the main European funded co-operation Programme in the field of higher Education is the TEMPUS Programme that supports the modernization of Higher Education and creates and area of co-operation in countries surrounding the EU (Western Balkans, Eastern Europe, Central Asia, North Africa and the Middle East). One of the main features of TEMPUS is the introduction of innovative teaching and learning methods through regional and international co-operation aimed to have an impact on higher education policies, making the different system in higher education more compatible and comparable.

New educational requirements and novel educational methods, such as Video Conference, supported by new telecommunication technologies enable almost instant access to latest educational materials and methodologies.

In this framework the TEMPUS Project VICES (Videoconferencing Educational Services), carried out by the University of Florence and the Ss Cyril and Methodius University in Skopje, was launched and financed by the European Commission for the period 2009-2012.

This project will introduce a new approach towards treatment of Information Communication Technologies at University level with the purpose to increase the virtual student and academic staff mobility. This approach will also enable higher level of harmonization of different curricula among partner institutions and at international level. This will increase the usage of new ICT technologies within the educational process, making it more efficient in the same time.

Acknowledgment

The authors acknowledge with gratitude the European Commission, DG Education and Culture who funded the VICES Project under the TEMPUS IV Programme.

References

1. Caporali, E.: How to design an Environmental and Resources Engineering Curriculum: The DEREC project experience. In: Caporali, E., Tuneski, A. (eds.) DEREC Development of Environmental and Resources Engineering Curriculum. Towards a new curriculum – The DEREC experience, pp. 9–18. Firenze University Press (2009)
2. Council of Europe. Convention for the Protection of human Rights and Fundamental Freedoms (September 2009), http://www.echr.coe.int/echr
3. Office of the high Commissioner for Human Rights. International Covenant on Economic, Social and Cultural Rights (September 2009), http://www.ohchr.org/EN/Pages/welcomePage.aspx
4. Towards the European Higher Education Area. Bologna Process. The official website 2007-2009 From London to Benelux and beyond (September 2009), http://www.ond.vlaanderen.be/hogeronderwijs/bologna/

5. Confederation of the EU Rectors' Conference and the Association of European Universities (CRE) The Bologna Declaration: an explanation (September 2009), http://ec.europa.eu/education/policies/educ/bologna/bologna.pdf
6. European Commission Education and Training Lifelong Learning Programme overview (September 2009), http://ec.europa.eu/education/lifelong-learning-programme/doc78_en.htm
7. European Commission Education and Training External Programmes and Policies Erasmus Mundus (September 2009), http://ec.europa.eu/education/external-relation-programmes/doc72_en.htm
8. The Tempus programme - Project overview (September 2009), http://ec.europa.eu/education/external-relation-programmes/doc70_en.htm
9. Videoconferencing Educational Services (September 2009), http://vices.marnet.net.mk
10. Borri, C., Guberti, E.: The contribution of Tempus Projects to mutual recognition of engineering study programmes across Europe. In: Caporali, E., Tuneski, A. (eds.) DEREC Development of Environmental and Resources Engineering Curriculum. Towards a new curriculum – The DEREC experience, pp. 21–25. Firenze University Press (2009)
11. Council of Europe. Convention on the Recognition of Qualifications concerning Higher Education in the European Region, Lisbon 11.IV.1997 (September 2009), http://conventions.coe.int/Treaty/en/Treaties/Html/165.htm
12. Bates, A.W., Bates, T.: Technology, E-learning and Distance Education. Rutledge Press (2005)
13. Eisenstadt, M., Vincent, T. (eds.): The Knowledge Web: Learning and Collaboration on the Net. Knowledge Media Institute. Kogan Page, London (2000)
14. Jeong, C., et al.: Context Aware Human Computer Interaction for Ubiquitous Learning. In: Proceedings of HCI, Beijing, China, pp. 364–373 (2007)
15. Surendar, C.: Lecture video capture for the masses. In: Proceedings of the 12th annual SIGCSE conference on Innovation and technology in computer science education (ITiCSE 2007), pp. 276–280 (2007)
16. Hutanu, A., et al.: Uncompressed HD video for collaborative teaching — an experiment. In: Proceedings of International Conference on Collaborative Computing: Networking, Applications and Work sharing, pp. 253–261 (2007)
17. Nishinaga, N., et al.: Enabling a cross-cultural collaborative community: networking technologies to form meaningful environments for higher education. In: Proceedings of the Fifth International Conference on Information Technology Based Higher Education and Training, ITHET 2004, pp. 203–208 (2004)
18. Dickson, P., et al.: First experiences with a classroom recording system. In: Proceedings of the 14th annual ACM SIGCSE conference on Innovation and technology in computer science education, pp. 298–302 (2008)
19. Hürst, W.: Indexing, searching, and skimming of multimedia documents containing recorded lectures and live presentations. In: Proceedings of the eleventh ACM international conference on Multimedia, pp. 450–451 (2003)
20. Hearnshaw, D.: Towards an Objective Approach to the Evaluation of Videoconferencing. Innovations in Education and Teaching International 37(3), 210–217 (2000)
21. Stephenson, E.J., et al.: Electronic delivery of lectures in the university environment: An empirical comparison of three delivery styles. Computers & Education 50(3), 640–651 (2008)

A New Solution for Workflow and Document Management Used for University Management in WODOMI Project

Thomas Biskup[1], Marjan Gusev[2], Gjorgji Manceski[3], and Pece Mitrevski[3]

[1] QuinScape GmbH, Dortmund, Germany
thomas.biskup@quinscape.de
[2] University "Ss. Cyril and Methodius", Skopje, Republic of Macedonia
marjan@ii.edu.mk
[3] University "St. Clement Ohridski", Bitola, Republic of Macedonia
{gjorgji.manceski,pece.mitrevski}@uklo.edu.mk

Abstract. In this paper we describe the architecture and the implementation of a new solution capable of building model-based web portals, used not only for university management, but with a much broader perspective. It is not just a replica of modern ERP systems, or portals that create content management solutions, but a novel solution proposed to be an efficient application for workflow and document management in SMEs.

Keywords: Portal, knowledge management, workflow management, document management.

1 Introduction

Enabling workflow and document management of all administration activities at the university and faculties was the main goal of the WODOMI project. Its wider objective was to establish a sustainable web-based network system, used as an interoperable environment for university workflow and document management, leading towards implementation of the "integrated University" concept and the goals recommended by the Bologna process. The development methodology used to set up the new solution is the Conceptual Programming approach (CP) described in details in [1]. Conceptual Programming consists of two major main methods – Conceptual Model Driven Software Development (CMDSD) and Concept Driven Project Management (CDPM) aspiring to develop individualized, evolvable and flexible solution in the context of small to medium sized enterprises (SMEs). The underlying philosophy of this approach is used to model workflows and to create a software tool that builds portals with integrated workflow solutions. The idea was extended towards use in integrated university environment, as a part of WODOMI project (modeling of workflows and processes required by integrated universities).

The efforts put in the "requirements analysis" and the "synthesis" part of the project, delivered development of a sophisticated design of workflows and document management, and finally resulted with a suitable modeling and portal creation tool with realization of integrated workflows. Comparing with existing Enterprise Resource Management solutions, the idea to create a portal that models and builds portals with integrated workflow solutions is new and not present in any of the ERP systems. ERP packages contain “embedded” business process knowledge. However, these packages are not able to represent business processes explicitly, in a natural way. Therefore, ERP packages (just as other legacy transaction processing applications) fail to provide tools for managing such business processes [2]. Most of the existing solutions are robust ERP systems and they present development environment with special language and programming support rather than the idea to have a relatively simple portal solution with modeling tool and engine that supports efficient run of the integrated workflow solution. Comparison with the existing solutions like Enterprise Architect [3], Intrexx [4], etc., shows that the approach used in these solutions does not integrate complete workflow solutions, but only portals with content management features. On the other hand, workflow management applications are often expected to solve this problem of business process management. However, integration of workflow management applications with data-centric applications is not very straightforward, especially when resources have to be shared in a business process. Workflow management is designed to focus on a single case (process instance), and all the resulting interference between cases (e.g. sharing resources) leads to problems. Therefore, an integration of workflow management with legacy transaction processing does not provide a satisfactory solution for integrated data and/or resource management [5]. This is the main motivation in the development of such a solution: the existing ERP systems are too expensive, robust and complex, while the portal creating content management systems do not integrate workflow solutions with all the features.

2 System Architecture

The solution is implemented as a web application. It is a development tool that models and creates web site starting point, as well as a structured presentation, so that customers can find a centralized starting place for access to consolidated enterprise-related functions. The package may be customized to varying degrees of enterprise or individual specificity. Portal software typically features a lot of complexity, automation, organization, and interactivity. The architecture of the solution is presented in Fig. 1. Logically, the application is divided in number of levels. Each level provides certain degree of abstraction and modularity (Fig. 3).

Database Provider is a system module that abstracts the work with the database and provides a simple programming interface. Its design is modular, for the purpose of providing easy access to different types of databases like SQL, Oracle, MySQL, etc. [6].

Base Object Model defines all of the objects/model/structures that will be used in the application. Also, the internal structure of the portal is defined.

Business Objects/ Logic is a system module where all of the business objects and logic in the application are defined.

Service is a system module that encapsulates some business process running in the application. Example: user login process, check for user rights, etc.

UI Process Components is a system module that represents precisely defined process by which the user input is accepted and its processing.

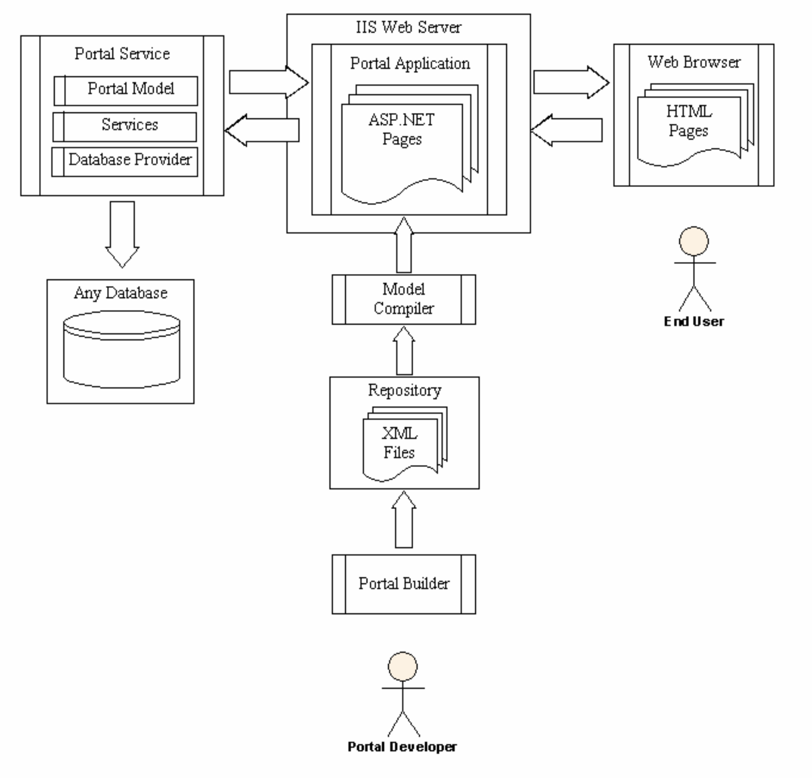


Fig. 1. Overall system architecture

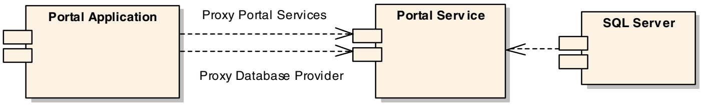


Fig. 2. Portal application deployment

Portal service is a background service for processing service requests and enables access to the database (Fig. 4). PortalServicesProxy provides availability of portal services. Portal Application represents the actual web application generated by the client application (Fig. 5). ClientServicesProxy enables logic for communication between clients and services in Portal Service. Proxy Database Provider provides mechanism for the client to access the Database Provider level of the Portal Service. Portal Builder is a client application (IDE) for developing web applications (Fig. 6).

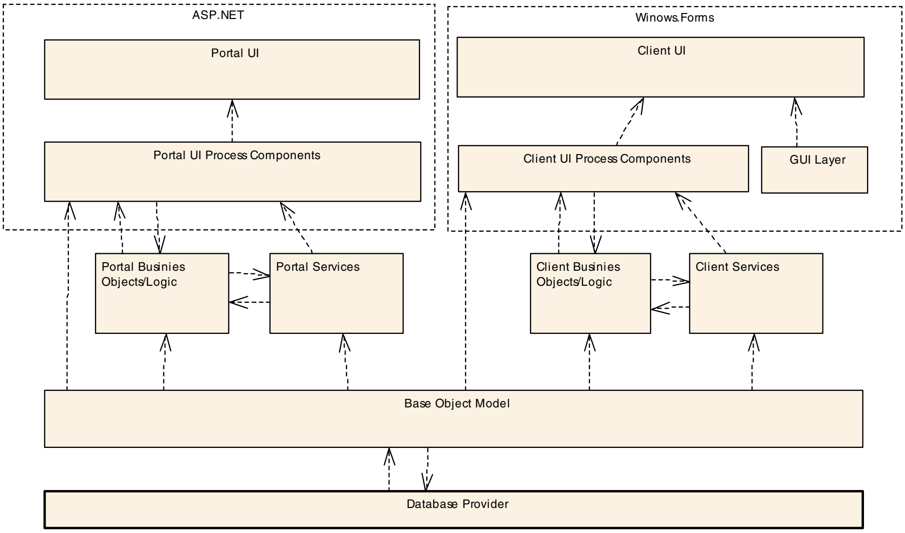
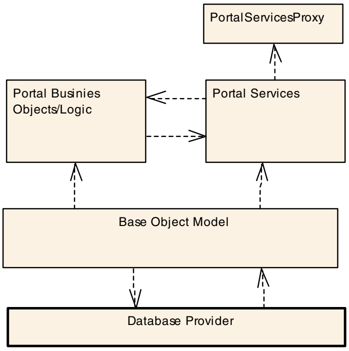

ASP.NET
Portal UI
Portal UI Process Components
Portal Businies Objects/Logic
Portal Services
Winows.Forms
Client UI
Client UI Process Components
GUI Layer
Client Businies Objects/Logic
Client Services
Base Object Model
Database Provider

Fig. 3. Application layers



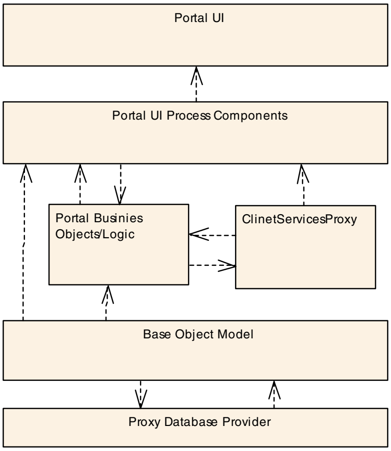


Fig. 4. Portal service layers

Fig. 5. Portal application layers

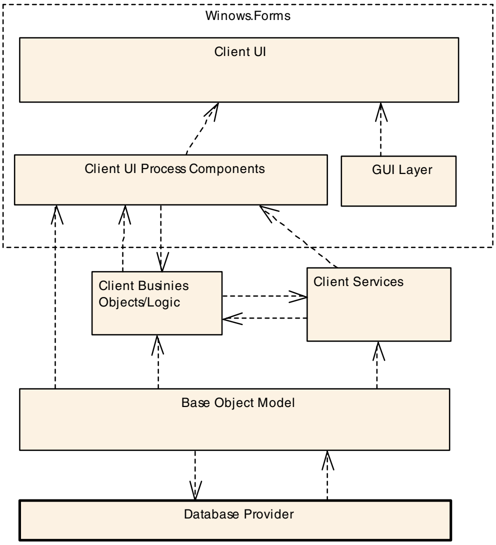


Fig. 6. Portal builder layers

2.1 Database Provider Layer

Database Provider level abstracts the logic for access to the database and enables efficient programming interface. All the semantics for physical access to the database is encapsulated, which is done through collection of Database Provider objects. Each Database Provider object is an abstraction and encapsulation for managing a certain type of business objects. Database Provider implements IDBProvider interface.

Database Provider Manager holds all the Data Providers available to the application and it is the only place for access to a certain Database Provider. The following types of providers are implemented: IDBAccountProvider - management of the user accounts of the portal; IDBPermissionProvider - management of the types of the permissions on the portal; IDBOrganizationProvider - management of the organizational model of the portal; IDBOrganizationUnitProvider - management of the organizational entireties of the portal (company); IDBRoleProvider - management of the roles in the portal; IDBTokenProvider - management of the security tokens of the portal; IDBDataModelProvider - management of the data model of the portal's internal structure.

2.2 Service Layer

System provides services for support for the other parts of the application. The service is a sum of help functions, properties, events which belong to one logical unit. Services are created once and are always available for all the parts of the application.

Service Manager holds all the services available to the application and is the only place for accessing some service. All requesting access to some service must be realized through IServiceManager. Service level provides following objects:

PortalServiceManager - services for support of the portal application and portal service; ClientServiceManager - services for support of the application for developing portal applications; PortalServicesProxy - makes the services from the Portal Service Manager available through .Net Remote; ClientServicesProxy - remote access to the services of the Portal Service Manager; it is used by the portal application when in need for some service.

2.3 Base Object Model

This level defines all the basic interfaces and objects of the portal. The model designer is used by the client. Model designer is provided by those models that are composed from more basic models, like: page, process etc.

Each model of the internal structure of the portal implements the following interfaces: IBaseModel is basic interface that is implemented by the rest of the models; IModelRepository is the object that transforms the model to XML file [7] and vice versa. IModelCompiler is object responsible for transforming the model in adequate script, in this case in ASP.NET [8]. Every model compiler contains ASP.NET template for the specific model; IModelPropertyEditor defines user interface for processing and adjusting the characteristics of the model; IModelDesigner defines user interface for designing the model. IRequestPermissions is interface that takes care about permissions/rights for access.

Portal model contains the name of the portal and references from all of the models that build the internal structure of the portal.

Organization model defines the organizational foundation of the company that owns the portal. Organizational model consists of several sub-models: IOrganization - describes the organization owning the portal; IOrganizationUnit - describes one organizational unit of the organization; IPerson - describes the persons in the organization.

UserAccount and Security model defines the users and their rights in the portal. It consists of the following models: IUserAccount - the module where each portal user gets user account consisting of unique username and password used for logging; IRole – the module that presents the user roles defined with certain rights (user can be a member of one or several roles); IPermission - Permissions defined in the portal (access, access to read, access to write, etc.); ISecurityToken - Every user, once successfully logged in the portal, gets security token. This token is kept in the session object of the ASP.NET.

2.4 Portal Data Model

Portal's data model consists of collection of data groups (Data Group), collection for the relations between data groups and collection of Views (Data View) for the data groups. IPortalDataModel unites the whole data model. IDataGroup is used for logical view of the physical table that would be created on the SQL server. IDGColumn – describes the column in the data group. IDGRow – describes one group of the data row. Row contains a collection of IDGRowItem. Number of IDGRowItem(s) is equal to the number of columns in the data group. IDGRowItem – maps column in the row of the data group. The portal's data group model is operated by adequate Data Provider from the Database Provider level.

2.5 Page and Control Model

Control is the basic UI element through which the user interacts with the application. For the needs of the portal the following controls are defined: IPortalControl – is the basic interface from which all of the other controls are inherited (supports and implements the IBaseModel interface and its characteristics); IActionControl - executes some transition from some process; IStaticControl - gets values in the process of designing the portal; IDataBindControl - enables connecting with a column of a data group from the portal's data model (two groups of this control: IEditControl - enables editing of the value of the column they map; and IViewControl – controls the preview of the value of the column they map). The portal page is actually a collection of controls. Each page supports and implements the IPortalControl interface. Each page implements IRequestPermission interface, where the rights that user must have to gain access to the page are defined.

2.6 Process Model

Process model supports modeling and implementing a business process in the application. Generally, a process is considered to be a graph of pages and transition between pages. Each node represents state of the business process that is modeled, and consists of a page and a transition. Page represents the user interface for a certain state of the business process. Transition represents an array of actions that are performed before jumping to the next node (state) from the business process.

2.7 Navigation Model

Generally speaking, the portal is a sum of pages, so one of the primary things that must be enabled is management of the navigation between pages. The navigational model represents a tree of navigational nodes, which describe the page that is loaded when a node is selected. Navigation saves and shows the hierarchical structure of the navigational model. It has the following properties: Name – name of the navigational model; and Nodes – collection of navigational nodes. INavigationNode represents one node of the navigational model and provides properties for building the hierarchical structure of the navigational model. INavigationNodeCollection represents a collection of navigational nodes.

2.8 Security Model

User Rights. User Rights are implemented using roles. Roles are kind of workgroups, and users can belong to one of more roles. Domain users (if they exist) are integrated in the solution, too. If a user is not defined in the solution , he/she is automatically treated as guest.

Page Security. During a process, it is necessary for every user to have the rights over a certain page, while at the same time the access for some other users is denied. Page right assignment is done by choosing Permissions.

Transition Security. For each transition, a list of users that can execute it might be given in the Process Designer.

2.9 Layout Model

Layout Model (a HTML template) provides positioning and placing of the models of the internal structure of the portal that make the GUI.

2.10 GUI Layer

GUI layer enables building of an IDE (Integrated Development Environment) for building web-portals (Fig. 7).

Workspace contains Views and Pads and does not know anything about the way they are shown on the screen. Views are, in fact, windows in the Workspace and have editable content. They hold the editors of the models. Pads are windows with tools. Workspace Layout is responsible for all the logic of showing Views and Pads on the screen and provides implementation of a Workspace Window where Views are shown. Workspace Window maps the basic functions of a window and contains View in itself.

IWorkspace maps the described model. IWindowContent consists of components that present editable contents. IDesignerModelContent is a special type of Window Content, which enables a model of the internal portal structure to be shown in the designer. IPadContent is window with tools that makes it easy for the user to finish some job. They are always active and might be visible or not. IWorkspaceWindow is a window used for showing the content from the IWindowContent components. Workspace Layout provides implementation of IWorkspaceWindow. ILayoutManager is responsible for the way Pads and Views are shown. It manages all the work about the user interface (selection of a specific window, showing and hiding of some Pad, creating a new window).

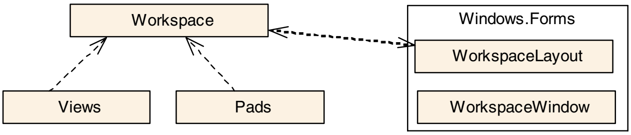


Fig. 7. Internal portal structure

2.11 Document Creator

This creator generates a report as a Word document. This is done by selecting the action on the given transition and setting it through its configuration dialog. For using Word documents in the solution, Word Action is implemented. For the purpose of this operation, a Word template has to be prepared previously. Everything that is put between tags is considered to be variable and is automatically read in the configuration dialog. In a document, the following tags appear: <<archive number>>, <<archive date>>, <<faculty name>>, etc. These variables represent default values that are filled automatically by the current Portal Page (i.e. from the current database). This kind of report gives opportunity for developing different documents without changing code in the Workflow Management Portals.

2.12 Multi-Language Support

Main purpose of the multi-language support is to give the user an opportunity to adjust the user interface according to the spoken language. It has the following features: 1) user interface is adjusted to accept different sets of characters and special characters from a certain language; 2) language files are not hard–coded and are easily added in the application. The idea behind the multi-language support is based on the requirement that each string that has to be translated in different languages should get a unique identifier (StringID). For each supported language there is a file with pairs of StringID and its representation.

Language definition is stored in an XML file (Languages.xml). For each supported language there is a different XML file languagename_lang.xml where pairs of StringID and correspondent representations are stored (also with the appropriate icon). Multi language support for IDE is a collection of precisely defined words defined during the development of the IDE application. Multi language support is provided by two services: ILanguageService – loads a language and provides access to specific words; ILanguageManager – used to add new languages into the system and to add new words in the respective language files.

2.13 Portal Validation

For each model included in the portal, there is a certain group of rules which must be respected so that it might be correct. Some of the rules for which validation is made are: check-up of all the links, whether there is transition connected to all action controls, whether there is a page in every node from the process, whether every page model has a unique ID, etc. It is necessary to make validation of the portal so that one can find defaults made during the portal design phase. As the portal test is impossible until it is sent to a web server, this is the only way to verify the correctness of the portal, according to the above rules.

2.14 Portal Deployment

This tool in the Portal Builder allows sending the portal to a web server where the Portal Engine has already been installed. Portal deployment is possible via internet.

3 Conclusion

The main contribution presented in this paper is the idea to create a portal builder tool with modeling features and integrated solutions for workflow and document management. This is rather a new idea, accomplished in the form of a small and affordable software solution, not present in any of the existing Enterprise Resource Management systems (they mostly refer to integrated development environments), nor in small portal building solutions (mainly concerned with content management features).

The new architecture, as well as the main features of the portal builder and portal compiler used to create portal applications with integrated workflow solutions, were presented in this paper. The portal builder software consists of a total of 2,112 files and 354,810 lines of code. By means of this tool, three workflow management portals were created for three different universities in the Republic of Macedonia (University "St. Clement Ohridski" – Bitola, University "Ss. Cyril and Methodius" – Skopje and the South East European University – Tetovo). In accordance with the WODOMI project goals, eight complex workflow processes have been covered at each university, along with a process for user administration.

References

1. Biskup, T.: Agile fachmodellgetriebene Softwareentwicklung für mittelständische ITProjekte. PhD thesis, Carl-von-Ossietzky University of Oldenburg (2009)
2. Szirbik, N., Wortmann, H.: Bridging the gap between ERP and WFM using agents. In: Proc. of the Intl. IMS Forum, Como, Italy (2004)
3. Enterprise Architect - Advanced Modeling & Design Platform - UML Tools For Business, Software and Real-time Systems, `http://www.sparxsystems.com/`
4. Intrexx - Enterprise Portal Software from the experts, `http://www.unitedplanet.com`
5. Cardoso, J., Bostrom, R.P., Sheth, A.: Workflow Management Systems vs. ERP Systems: Differences, Commonalities, and Applications. Information Technology and Management 5, 319–338 (2004)
6. Patrick, J.J.: SQL Fundamentals (2002)
7. Elliotte, R.H., Means, W.S.: XML in a Nutshell. O'Reilly Media, Sebastopol (2004)
8. Khosravi, S.: Professional ASP.NET 2.0 Server Control and Component Development. Wiley Publishing, Chichester (2006)

Author Index

Aalmink, Jan 257, 327
Ackovska, Nevena 85
Amer, Mahmoud 227
Andonovic, Ivan 21
Andovska, Mirjana 75
Andrade, Rafael A. Espin 35
Angeleski, Marjan 265
Apostolski, Vladimir 237
Armayor, Dania Pérez 285
Arsov, Ivica 315

Bakeva, Verica 377
Bakraceski, Goran 237
Balloul, Lama 257
Banerjee, Nilanjana 33
Batista, José Antonio Díaz 285
Biskup, Thomas 451
Bozinovski, Stevo 85

Caporali, Enrica 439
Celakovski, Sasko 337
Chorbev, Ivan 75
Cingovska, Ivana 125
Cus, Franc 429

Davcev, Danco 115, 125, 135, 145, 337
de Castro, Roberto Pérez López 285
Descoteaux, M. 345
Dimitrova, Nevenka 33
Dimov, Zoran 115

Filiposka, Sonja 197

Gecevska, Valentina 429
Gilroy, Michael 21
Glagau, Jan 257
Gligoroski, Danilo 391
Goh, Hock Guan 21
Gómez, Jorge Marx 1, 35, 217, 227, 257, 285, 295
Gusev, Marjan 157, 275, 451

Haak, Liane 327
Hadzi-Velkov, Zoran 177
Hicks, James 33

Ivanoska, Ilinka 145
Ivanovska, Ilinka 135

Jakimovski, Boro 401, 411
Janeska, Margarita 265
Janevski, Angel 33
Janković, Dragan 247
Jordanovski, Ljupco 411
Jovanov, Mile 275
Jovanova, Blagica 305, 315

Kalajdziski, Slobodan 115, 125, 135
Kamalakaran, Sitharthan 33
Kirkova-Naskova, Anastazija 237
Klima, Vlastimil 391
Kocarev, Ljupco 167, 187
Koceska, Natasa 105
Koceski, Saso 105
Kojchev, Risto 167
Kon-Popovska, Margita 421
Kotevski, Zoran 357
Kulakov, Andrea 55, 95
Kwong, Kae Hsiang 21

Lameski, Petre 95
Lucito, Robert 33

Mahmoud, Tariq 295
Manceski, Gjorgji 451
Marković, Ivica 247
Markovski, Smile 367, 377
Martinovska, Cveta 45
Memari, Ammar 227
Michie, Craig 21
Mileva, Aleksandra 367
Mirceva, Georgina 115, 135
Mirchev, Miroslav 187
Misev, Anastas 411, 421
Mishkovski, Igor 167
Mitreski, Kosta 207
Mitrevski, Pece 265, 357, 451
Mosquera, Inty Sáez 285

Naumoski, Andreja 207

Pepik, Bojan 135
Pérez, Alberto Morell 217
Pešić, Srebrenko 247
Peters, Dirk 327
Popovska-Mitrovikj, Aleksandra 377
Poupon, C. 345
Prčkovska, V. 345
Preda, Marius 305, 315
Preteux, Françoise 305, 315

Risquet, Carlos Pérez 217

Sahpaski, Dragan 401
Sasloglou, Konstantinos 21
Spasov, Dejan 157
Stojkoska, Biljana 145
Stosic, Jovan 177

ter Haar Romeny, B.M. 345
Trajanov, Dimitar 167, 197, 237
Trajkovik, Vladimir 439
Trajkovski, Igor 65
Trivodaliev, Kire 125, 135

Valdés, Ariel Racet 35
Valdiserri, Juna 439
Varadan, Vinay 33
Vasileska, Dragica 5
Velinov, Goran 401
Vilanova, A. 345
von der Dovenmühle, Timo 295

Wu, Tsungta 21

Zdravevski, Eftim 55
Zuperl, Uros 429